

AN EXPERIMENTAL EVALUATION OF SEVEN COMPOSITION SCALES

BY
DOROTHY SPEER

A DISSERTATION

submitted to the Board of University Studies of The Johns Hopkins University in conformity
with the requirements for the degree of Doctor of Philosophy.

JUNE, 1929

BALTIMORE
THE JOHNS HOPKINS PRESS
1929

AN EXPERIMENTAL EVALUATION OF SEVEN
COMPOSITION SCALES



Digitized by the Internet Archive
in 2026

AN EXPERIMENTAL EVALUATION OF SEVEN COMPOSITION SCALES

BY
DOROTHY SPEER

A DISSERTATION

submitted to the Board of University Studies of The Johns Hopkins University in conformity
with the requirements for the degree of Doctor of Philosophy.

JUNE, 1929

BALTIMORE
THE JOHNS HOPKINS PRESS
1929

COPYRIGHT 1929 BY
THE JOHNS HOPKINS PRESS,
BALTIMORE, MD.



The Lord Baltimore Press
BALTIMORE, MD., U. S. A.

ACKNOWLEDGMENTS

Without the co-operation of the persons to whom I herein acknowledge my deepest gratitude and indebtedness, the experiment would have been impossible.

To Professor Florence Eilau Bamberger I wish to express my sincere appreciation for the specific service, untiring efforts, and great encouragement given in the inception and direction of this study. In the statistical treatment of the investigation invaluable aid was rendered by Dr. Fowler Dell Brooks, for which I am deeply grateful. For many significant suggestions and ever-ready co-operation in the prosecution of this study I am indebted to Professor Edward F. Buchner.

Valuable assistance in the collection of data was given by numerous principals, teachers, and the Baltimore Bureau of Research. I express my thanks to all these, especially to Dr. John Stenquist, Miss Grace Kramer, and Dr. Harold Chapman.

I wish to acknowledge also the exceptional professional spirit manifested by Dr. Ellwood P. Cubberly and Miss Dorothy Putnam of Stanford University, and Dr. C. C. Certain, editor of the *Elementary English Review*, in their co-operation over a period of two years in securing answers to questionnaires employed in this study.

Several friends and relatives who gave unsparingly of their time and knowledge in the rating of themes deserve special appreciation. These are Miss Eunice Crabtree, Miss Mary Braun, Miss Lillian Braun, Mrs. Florence Huether, Miss Frieda Speer, and Mr. Herbert Speer.

For unlimited and indispensable help in the statistical calculations, records of long tables, clerical management, and valuable suggestions, I am profoundly grateful to my cousin, Hilda Speer.

D. S.

CONTENTS

CHAPTER	PAGE
Introduction	I
I. Problem and Materials.....	3
A. Problem	3
B. Presentation of Data.....	8
1. Scales Used	8
2. Collection of Compositions.....	16
II. Technique of the Experiment.....	20
A. Selection of Compositions.....	20
B. Preparation of Compositions	24
C. Selection of Judges.....	25
D. Directions to Judges.....	25
E. Practice Effect	27
F. Training of Judges.....	27
G. Subjective Ranking of Scales by Judges in Order of Preference.....	30
III. Criteria Used in This Study.....	31
IV. Results of the Experiment.....	34
A. Objectivity of Scales as Determined by Criterion I	34
B. Objectivity of Scales as Determined by Criterion II	44
C. Objectivity of Scales as Determined by Criterion III	46
D. Objectivity of Discourse Distinction Scales versus Non-Discourse Distinction Scales	47
E. General Merit versus Diagnostic Scales....	50
F. Effect of Training on Hillegas and Non- Hillegas Derivatives	51
G. Effect of Special Factors of Scales on Their Objectivity	52
1. Number of Scale-Steps.....	52
2. Statement of Merits and Defects....	53
3. Number of Samples per Scale-Step..	54

CHAPTER	PAGE
H. Special Adaptability of the Lewis Scales for Letters	55
I. Subjective Ranking of Scales by Judges in Order of Preference.....	56
J. Administrative Feasibility	58
1. Time	58
2. Grade Suitability	59
K. Use of Scales in School Surveys.....	60
V. Summary of Conclusions.....	63
A. Specific Conclusions	63
B. Scale Suggestions to English Teachers and Supervisors	67
C. Further Problems	69
APPENDIX	
I. Evidence on the Inapplicability of the Harvard-Newton English Composition Scales to the Sixth Grade.....	71
II. Evidence of Improvement Resulting from Training—Criterion I.....	73
III. Bibliography	74
IV. Detailed Instructions to Judges.....	80
Vita	86

FIGURES

FIGURE		PAGE
1.	Hudelson English Coefficients of Variation.....	38
2.	Hudelson Typical Coefficients of Variation.....	38
3.	Leonard Coefficients of Variation.....	39
4.	Lewis Coefficients of Variation.....	39
5.	Nassau Coefficients of Variation.....	39
6.	Van Wagenen Coefficients of Variation.....	40
7.	Willing—Story Value—Coefficients of Variation.....	40
8.	Willing—Form Value—Coefficients of Variation.....	40
9.	Quality of Fifty Specimens Chosen to Study Improvement Due to Training.....	41
10.	Quality of Forty Specimens Chosen to Build up Criterion II.	44
11.	Harvard-Newton Coefficients of Variation.....	72

TABLES

TABLE	PAGE
1. Distribution of Composition Topics.....	21
2. Age-Grade Progress Table for 6B White Pupils in Baltimore as of February, 1927.....	22
3. Age Distribution with Respect to Composition Topic Choice	23
4. Intelligence Distribution with Respect to Composition Topic Choice	24
5. Mean, Standard Deviation, and Coefficient of Variation of 100 Compositions Grouped by Type of Discourse for Each Scale	35
6. Average Coefficient of Variation of Each of Seven Composition Scales	36
7. Mean Coefficients of Variation and Standard Deviations Corresponding to Figures 1-8.....	37
8. Improvement of 50 Selected Compositions Representing Five Qualities and Their Corresponding Improvement in Terms of Coefficient of Variation.....	42
9. Improvement of 50 Selected Compositions Representing Five Improvements and Their Corresponding Qualities in Each of Seven Scales.....	43
10. Building up Criterion II.....	45
11. Advantage in Objectivity of Marks Assigned by Use of Composition Scales over Chance.....	46
12. Arithmetic Average of the Means, Standard Deviations, and Coefficients of Variation for Five Groups of Compositions Exemplifying Different Types of Discourse...	48
13. Average Coefficients of Variation of Discourse and Non-Discourse Distinction Scales.....	49
14. Average Coefficients of Variation of Diagnostic and General Merit Scales.....	51
15. Average Coefficient of Variation of Hillegas and Non-Hillegas Derivatives	52
16. Average Coefficients of Variation of "Short," "Average," and "Long" Scales.....	53
17. Average Coefficients of Variation of Scales with and without a Statement of Merits and Defects Following Each Scale Sample	54
18. Average Coefficients of Variation for Scales Containing One Sample per Scale-Step (Group I) and Scales Containing More Than One Sample per Scale-Step (Group II)	55
19. Average Coefficient of Variation of Letter Compositions on Lewis <i>vs.</i> Other Scales.....	55
20. Ranks of Scales in Order of Judges' Preference.....	56
21. Average Time per Judge per Composition.....	58
22. Undistributed Scores After Training.....	60
23. Scales Ranked According to Criteria I, II, III.....	63
24. Harvard-Newton English Composition Scales.....	71
25. Improvement as a Result of Training in Terms of Coefficients of Variation.....	73

INTRODUCTION

To state that one of the most outstanding and far reaching of modern educational movements is that of standardized measurements in all phases of school activity is a mere platitude. The most superficial inspection of the literature will quickly reveal this. Even composition, perhaps the most complex and elusive of all subjects to measure, has not escaped this all-embracing inquisition. Just because of the difficult nature of the problem, little has been agreed to by common consent regarding the objectivity and general value of the available instruments—the composition scales. An examination of the references composing the bibliography appended to this study exposes the controversial atmosphere which envelops this subject.

The widespread dissatisfaction with the *status quo* of achievement in composition in all grades from the third through the secondary school, and even college, and the vagueness of standards have driven some to seek refuge in such scales as we have, and have forced others into a position of despair regarding the utter impossibility of extracting order from the chaos. Again, a perusal of the articles and texts listed in the bibliography will convince the reader of this condition.

Such scales as have been employed for various and sundry reasons by persons conducting surveys, bureaus of research, school supervisors and administrators, and progressive classroom teachers, have, in almost every case, been selected on purely arbitrary grounds. Evidence for this is expressed by Miss Johnston in the following quotation:¹ “In regard to stating any reason for their choice of a certain test, the surveys, as a rule, give little information beyond the fact that the

¹ Johnston, Nell Bomar. “The Use of Standardized Educational Tests in School Surveys.” *Educational Administration and Supervision*, Vol. 2, 1925, p. 602.

test has been used in surveys, and is desirable for purposes of comparison." The writer discovered exactly the same fact in her review of school surveys from 1924-1928, inclusive. The existence of these conditions, and the absence, so far as the writer is aware, of any thoroughgoing, complete experimental scrutiny of these measuring rods (that is, subjecting them simultaneously to the same set of experimental conditions), has led to this investigation, in which she hopes to throw some light upon the objective character, the administrative feasibility, the specific suitability, and other aspects of the nature of these scales.

CHAPTER I

PROBLEM AND MATERIALS

Problem. As the title states, this investigation is an experimental evaluation of seven composition scales. A more detailed statement of it would be: a study to ascertain experimentally (a) the objectivity of these scales and the improvement in objectivity as a result of a limited amount of training; and (b) their administrative feasibility. Under these two phases the writer will endeavor to answer the following questions:

OBJECTIVITY AND THE EFFECT OF TRAINING

1. How do the scales rank in objectivity when they are used by untrained judges? Which scale is the most objective in the hands of untrained judges? Which is the least objective?

2. How do the scales rank in objectivity when they are used by trained judges? Which scale is the most objective in the hands of trained judges? Which is the least objective?

3. On which scale is there the most improvement in objectivity as a result of training?

4. Upon which group of compositions—"Superior," "Good," "Average," "Fair," "Inferior"—is there greater gain in objectivity as a result of training? Therefore, for which type of compositions must the marker receive more training in order to grade objectively?

5. To what extent do scales which make discourse distinctions measure a given type of composition more objectively than non-discourse distinction scales?

6. Do scales which are modifications of the Hillegas Scale (hereinafter referred to as "Hillegas derivatives") improve more in objectivity from the type of training we gave them than non-Hillegas derivatives? Hence, what type of material is suitable for practice before using the respective scales?

7. Which type of scales, general merit or diagnostic, require more training to make them objective in their use? Which is the best diagnostic scale, and which is the best general merit scale?

8. What is the effect of the absence of a standard for marking such a matter as form?

9. Is there less variation in judgment in scales which have few, coarsely divided scale-steps than in those which have more, finely divided scale-steps? What is the effect of training on the objectivity of each of these two types?

10. To what extent do the subjective rankings of the scales in order of merit by the judges agree with the experimental determination of the relative objectivity?

11. Does a statement of the analysis of the merits and defects following each sample on a composition scale make for greater uniformity of judgment?

12. Do scales which have more than one sample per scale-step make for greater objectivity than those which do not have such?

ADMINISTRATIVE FEASIBILITY

1. Which scale gives the most objective results in the least time before training? After training?

2. Have educational experts who conducted surveys always used the most objective and feasible scales?

3. Which of the scales seems to fit these sixth grade compositions best?

It will be noticed that an answer to these questions will throw considerable light also upon the desirable physical make-up of a scale, which may be helpful in serving as guides to scale constructors.

Professor Earl Hudelson² says, "Composition is, to be sure, a highly subjective educational product; yet it conforms to the general principle that whatever exists, exists in some amount, and that any perceptible difference can be measured. The more the subjective aspect of composition can be stripped of its formal, impersonal elements, and these elements measured objectively, the less chance there will be for variation in teachers' judgments." To determine just how objective the scales we now have are, seemed important in the light of this statement.

In the same article Hudelson² lists the aims and purposes of composition scales. These are: "(1) not to improve writing ability; (2) to test impartially the various methods of teaching composition by measuring their results; (3) to measure those results in accurate, objective, stable, and understandable terms; (4) to furnish a common basis for comparing the writing proficiency of different pupils within the same class or school, or that of pupils in different classes or schools; (5) to fairly classify pupils in composition; (6) to

² Hudelson, Earl. "The Development and Comparative Values of Composition Scales." *English Journal*, Vol. 12, March, 1923, p. 164.

grade them justly within the group; (7) to enable teachers to discover their reliability in judging the merit of English composition; (8) to furnish pupils an incentive to self-competition." Such a formulation of aims emphasizes the need of such an investigation as the one which the writer herein proposes.

The extent to which composition scales have been used in actual school room situations aside from experimentation for pure research adds further proof for the urgent need for an answer to the questions raised by the writer in her statement of the problem of this study. More than half the references composing the bibliography appended to this investigation reveal just such a demand. It is not the experimenter's purpose to review all of these, but she will take time to mention just one or two representative cases.

The teachers' colleges in Minnesota attempted to answer, by use of Hudelson's Maximal Composition Ability Scale, this problem:³ "This was an experiment upon examinations which were to determine what students should be asked to enter a non-credit course, and what students should take the regular composition work of the college."

M. D. Brown, and M. E. Haggerty⁴ attempted to measure the improvement in English composition by rating the composition work of seventy-eight high school students throughout a period of twelve weeks to answer the following questions (by means of the Harvard-Newton Scales):

1. Do students who write a weekly theme under the ordinary school conditions improve in ability to write?
2. Is improvement characteristic of all or only a portion of such students?
3. What are satisfactory limits of improvement for high school students?
4. What are the necessary conditions for satisfactory improvement?
5. What hindering conditions may be avoided?

³ Barden, Carrie. "More Experiments with Scales." *English Journal*, Vol. 13, November, 1924, p. 649.

⁴ Brown, M. D., and Haggerty, M. E. "The Measurement of Improvement in English Composition." *English Journal*, Vol. 6, October, 1917, pp. 515-527.

The question arises: To what degree is the Harvard-Newton Scale suited for such use? Why did the experimenters choose this scale?

In order to ascertain the extent to which teachers all over the country use the available instruments for measuring composition, and what their preferences and their practices are in such use, the writer devised and sent out to educational authorities over the whole United States the questionnaire which is reproduced in this study at the end of this chapter. In brief, the answers to this furnish information concerning what scales are commonly used; where they are used; which are the preferred scales and why; what is the value of the scales as judged by the users; how frequently the scales are employed by the same persons, and to what topics raters apply these scales. M. J. Van Wagenen⁵ has said, "Although scales for measuring general merit in English composition have now been in existence for several years, it is astonishing how few teachers of English make use of them or even know about them." This was exactly the writer's experience when she received returns from the questionnaires. In spite of the fact that these were distributed at such conventions as those of the National Council of Teachers of English both in Chicago and Baltimore, that the questionnaire was published in the "Elementary English Review," and that numerous heads of Schools of Education, elementary and secondary school supervisors, co-operated in distributing and collecting these, only forty-one replies from twelve states in the Union were received. These states were: Arizona, California, Colorado, Illinois, Indiana, Maryland, Michigan, New York, Oregon, Pennsylvania, Washington, and West Virginia. Practically all the sections of our country are represented. A summary of results follows. On such slight evidence no definite conclusions can be drawn, but the returns are, at least, interesting and probably symptomatic of the national state of affairs in regard to composition scales.

⁵ Van Wagenen, M. J. "The Accuracy with Which English Themes May Be Graded with the Use of the English Composition Scales." *School and Society*, Vol. 11, April 10, 1920, p. 441.

SUMMARY OF QUESTIONNAIRE REPLIES

Replies received, 41; number using scales, 21; number not using scales, 20.

The scales which were used by the forty-two persons⁶ who replied are:

Hillegas, 1; Harvard-Newton, 5; Hudelson English, 6; Hudelson Typical, 9; Leonard, 1; Lewis Letter, 1; Nassau Supplement, 8; Van Wagenen, 3; Willing, 6; Breed and Frostic, 1; Hudelson Maximal, 1.

The scales which were preferred and the number of users who preferred them are:

Hudelson Typical, 8; Lewis, 1; Nassau, 2; Harvard-Newton, 2; Hudelson English, 2.

The writer considers this list of preferences unrepresentative because they were limited by familiarity with two or fewer scales. This limited acquaintance was expressed by quite a number of the persons who replied to the questionnaire.

The values which the thirty-four persons replying assigned to the use of composition scales are:

Objectivity, 19; diagnosis, 6; pupil self-rating, 3; classification of pupils, 1; doubtful value, 3; no value at all, 2.

The frequency of the application of a scale during the scholastic year is:

Twice a term, 4; four times a term, 1; periodic class examination, 1; "frequently," 1; "seldom," 1.

Briggs⁷ claims that "to know what growth there has been, measurement by scale should be made in September and then not again until June, or perhaps the end of the half-year." The returns we received seem to point in the same direction. The application of the scale to compositions writ-

⁶ Totals for each item will be different from the number of questionnaires received because of the varied practice in answering the different items.

⁷ Briggs, Thomas H. "English Composition Scales in Use." Teachers College Record, Vol. 23, No. 5, November, 1922, p. 429.

ten on the same or a different subject from that in the samples on the scale employed is:

Same subject, 7; any subject at all, 5; varied usage, 3.

Some authors⁸ of scales have held that a composition scale should be applied only to compositions which treat of the same topic as those composing the scale which is being used. To judge from our replies, camps are about equally divided in this regard. It seems to the writer that the necessity of applying scales to only such themes as deal with the same topics as the scale samples would restrict the use of these scales considerably, and might even lead to coaching on the scale topics. Further, it might cause pupils who were allowed to see the scales to use the latter as models, and, thus, to lead them to imitate. This would vitiate the purpose of the composition scale.

In his dissertation Dr. Chapman⁹ reveals the fact that, out of a total of fifty-two bureaus of research in the United States, only three used composition scales at all. This took place during the academic year 1924-1925. The scale used was the Nassau County Supplement to the Hillegas Scale.

Evidence to justify the proposed investigation is thus seen to be plentiful, and answers to the questions raised in Chapter I will aid considerably in substituting definite knowledge for conjecture.

PRESENTATION OF DATA

Scales Used

The seven composition scales which were subjected to experimentation are as follows:

The Hudelson English Composition Scale (sometimes known as "The Virginia Supplement to the Hillegas Scale"), by Hudelson.

⁸Willing, M. H. "Scale for Measuring Written Composition." Public School Publishing Company (see "Directions for Using the Willing Scale for Measuring Written Composition").

Hudelson, Earl. "The Hudelson Typical Composition Ability Scale." Teachers' Handbook, Public School Publishing Co., Bloomington, Ill., p. 4.

⁹Chapman, Harold B. "Organized Research in Education." Ohio State University Studies, Ohio State Press, Columbus, Ohio, 1927.

The Hudelson Typical Composition Ability Scale, by the same author.

A scale for the Judgment of Composition Quality Only, by Sterling A. Leonard.

Scales for Measuring Special Types of English Composition, by Eugene E. Lewis.

Nassau County Supplement to the Hillegas Scale, by M. R. Trabue.

Van Wagenen English Composition Scales (often referred to as the "Minnesota English Composition Scales"), by M. J. Van Wagenen.

Scale for Measuring Written Composition, by M. H. Willing.

Two other scales which entered into the experiment were the Harvard-Newton Composition Scales, by F. W. Ballou, and the Thorndike Extension of the Hillegas Scale, by E. L. Thorndike. While the writer is well aware of the fact that the Harvard-Newton Scale is intended as an eighth grade composition scale,¹⁰ and that it would, therefore, be unsuitable in this experiment, which considers only compositions written by sixth grade pupils, it seemed desirable to find out whether at least the lower range of the scale might apply to sixth grade subjects. Furthermore, this scale has been used for the sixth grade in Brookline, Massachusetts, and for the seventh grade in Bloomington, Indiana, and even in grades four to twelve in the Port Townsend, Washington, Surveys. However, the evidence which we have proves conclusively the utter inapplicability of this scale to compositions of the sixth grade. This evidence is presented in Appendix I.

The Thorndike Extension of the Hillegas Scale was used for training the judges. A full explanation of the procedure followed in training the judges will be given later. It is necessary at this point to state only our reason for choosing this scale for practice. This was the fact that the best and largest collection of standard samples of English composition of known value have been gathered together by Thorndike and published in a little volume entitled, "English Composition, 150 Specimens Arranged for Use in Psychological and Educational Experiments." The complete reference is listed in the bibliography. These specimens have been stand-

¹⁰ Ballou, Frank W. "Scales for the Measurement of English Compositions." Harvard-Newton Bulletins No. II, September, 1912, p. 8.

ardized on the Thorndike Extension of the Hillegas Scale. A further reason guiding our choice was the fact that four of the seven scales are Hillegas derivatives (*i. e.*, are either modifications of this scale, or are evaluated in Hillegas terms), and, hence, the training given could be expected to transfer rather completely to the increased accuracy in the use of these scales. No very extensive material is available for practice in non-Hillegas derivatives. Such as does exist was utilized to ascertain the effect of training, which will be explained later.

Other scales which exist for the measurement of general merit in composition, but which were not included in this experiment are:

A Scale for the Measurement of Quality in English Composition for Young People, by M. B. Hillegas.

A Scale for Measuring the General Merit of English Composition in the Sixth Grade, by F. S. Breed, and F. W. Frostic.

Maximal Composition Ability Scale, by E. Hudelson.

The experimenter was very desirous of including the last two in this investigation, but was unable to procure them because they were out of print at the time of the beginning of the solution of this problem.

If the reader wishes to get a comprehensive view of the history of the making of composition scales, their nature, the arguments for and against their use, their purpose, and a general account of the status of the question past and present, perhaps the best single reference is the review of the subject given by Earl Hudelson in the Twenty-Second Yearbook of the National Society for the Scientific Study of Education, Part I, entitled, "English Composition, its Aims, Methods, and Measurement." (For full reference, see the bibliography number 31(a).) Only a very brief summary of the chief characteristics of each scale will be given at this point in order to serve as a basis for understanding the kind of investigation herein undertaken, and the reason for raising the questions stated under "Problem" in the beginning of this chapter.

Chronologically arranged, the nine scales which figure in this experiment are: the Harvard-Newton, the Thorndike,

the Willing, the Nassau, the Van Wagenen, the Hudelson English, the Lewis, the Hudelson Typical, and the Leonard.

The first of these has already been described sufficiently for our purpose. One more feature may be mentioned. It consists of four scales: one for narration, one for description, one for exposition, and one for argumentation. This we have been pleased to designate as a "discourse distinction scale." Each of the scales is composed of six samples rated A, B, C, D, E, F, respectively, with their corresponding values expressed in per cents, which differ for each letter from scale to scale. Each sample is followed by a statement of its merits and defects, which statement justifies the position which the given sample occupies in the scale as a whole. It is devised to measure the general merit of compositions, although it has been claimed that it may be used also for diagnostic purposes.

The Thorndike Extension of the Hillegas Scale is likewise a general merit scale. It does not make any discourse distinctions, nor is each sample in it followed by a statement of its relative merits and defects. Its steps are irregular, ranging from zero quality to quality 95. All in all there are fifteen scale-steps. Throughout the middle section of the scale several samples are given on each scale-step. Certain writers hold that this is a distinct advantage.¹¹

The Willing Scale is ostensibly a diagnostic scale. It emerged as an outgrowth of the Denver Survey, and is not, in any sense, a modification of the original Hillegas Scale. We, therefore, designate it as a "non-Hillegas derivative" in this study. Its outstanding characteristic is its provision for separate ratings for "story value" and "form value." The rater is to read each theme through first for its story value only, and is to determine the quality of this by comparison successively with the samples given in the scale, which are only eight in number, ranging in numerical value from 20 to 90. Having determined this, the rater is to count the number of errors in form (grammar, punctuation, capitalization, and spelling) per hundred words, and is to assign a

¹¹ Klapper, Paul. "Teaching English in Elementary and Junior High Schools." D. Appleton & Company, 1925, p. 190.

grade accordingly. The author gives no method for combining the two scores which are in terms of totally different units. The scale is intended for measurement of grades four through eight. Its scale-steps are few and coarse.

The Nassau Supplement to the Hillegas Scale is, as its name indicates, a "Hillegas derivative." It is a general merit scale which consists of only ten irregular scale-steps, ranging in numerical value from zero through nine. The samples on each step (and there is just one to a step) are short, and in no way commented upon in the scale. The instrument is claimed to be applicable from the fourth grade through the senior high school. It makes no discourse distinctions, nor does it make any attempt at diagnosis. This scale is a by-product of the Nassau County Survey. A set of twenty-five standard samples of known value are issued by S. A. Courtis which can be used for practice material in the use of this scale. By means of these a judge of composition can train himself to a known degree of accuracy before applying the scale to specific situations if he so desires.

The Van Wagenen English Composition Scales are both diagnostic and discourse distinction scales. They are non-Hillegas derivatives. There is one scale for each of three discourses: exposition, narration, and description, respectively. These are divided into fifteen, fourteen, and fourteen irregular scale-steps, respectively. For purposes of diagnosis, each sample is analyzed into three aspects, for each of which a separate grade is assigned. These aspects are thought content, structure, and mechanics. The three qualities are not evaluated in equivalent terms in the same scale, but each quality in each scale furnishes practically an equivalent scale for the same quality in either of the other two discourses. Thus, a 72 in thought content is not equivalent to a 72 in either structure or mechanics within the same scale or in either of the other two scales; but a 72 in thought content on any one of the scales is practically equivalent to a 72 in thought content on either of the other scales. Thought content values range from 6 to 100 on the exposition scale, 8 to 94 on the narration scale, and 8 to 91 on the description

scale. The corresponding values for structure range from 3 to 102, 4 to 94, and 9 to 104, on the respective scales; while the values for mechanics range from 1 to 97, 0 to 92, and 6 to 94, respectively. The samples are not followed by any sort of statement of merits and defects, nor is more than one sample given per scale-step. In the author's explanation on "How to Rate Compositions"¹² he gives a careful and detailed enumeration of the component elements which the rater is to consider under "Thought Content in Exposition," "Thought Content in Narration," "Thought Content in Description," "Sentence and Paragraph Structure," and "Mechanical Errors." In the same section Van Wagenen sets forth the combination of the three scores given each composition into a composite grade designated as "general merit." Unlike Willing, Van Wagenen has determined statistically the relative weight which should be given to each of these three elements: thought content, structure, and mechanics. He weights the first four, the second two, and the last one. Then he adds the result and divides by seven. The average is the general merit composite score. This last feature was provided to satisfy the universal demand on the part of teachers and administrators for one composite grade on a composition. Each of these three scales is equipped with a number of specimens for practice in rating themes. There are sixteen such for the exposition and the description scales each, and fifteen for the narration scale. There is a key to the real value of each of these themes, and a rater can ascertain his accuracy in rating compositions by this scale. The writer's personal experience is that these series are much too short.

In the Hudelson English Composition Scale we meet for the first time a scale which has equal step intervals. Like the Nassau, it is the result of a survey (made in Virginia, whence the name, "Virginia Supplement to the Hillegas Scale"), and is a general merit scale of the "Hillegas derivative" variety. It consists of sixteen samples, one to a scale-

¹² Van Wagenen English Composition Scales, M. J. Van Wagenen, World Book Company, New York (Yonkers), p. 2.

step, and has no statement of the relative merits and defects of its respective composition samples. The values of these steps range from 2 to 9.5. The scale is accompanied by three practice series of ten compositions of known value each, making a total of thirty practice compositions upon which the rater may train and determine his exact deviation from expert judgment.

The Lewis English Composition Scales represent the first attempt to measure specific types of writing, such as letters. Lewis has a separate scale for measuring (1) the quality of order letters, (2) letters of application, (3) narrative social letters, (4) expository social letters, and (5) simple narration. In the present study only three of these scales were utilized: those for order letters, narrative social letters, and expository social letters. The scale for order letters consists of nine, coarsely divided scale steps which range from 15 to 85 in value. That for narrative social letters consists of eight steps which range from 15 to 85; and the scale for expository social letters contains nine steps ranging from 5 to 80 in value. These samples are all evaluated in Hillegas terms and, hence, can be called Hillegas derivatives. Like most of the other scales, these have only one sample per scale-step, and are not followed by any statement evaluating their relative worth. The values progress irregularly from step to step. The first two of the letter scales are calculated to measure two very common types of business letters, while the last two measure two equally common types of friendly letters. All of the scales attempt to determine the general merit of the compositions, and are not diagnostic in any way.

The Hudelson Typical Composition Ability Scale is a novel effort to ascertain the average typical performance of pupils. It consists of eighteen samples evaluated in Hillegas terms and distributed over ten scale-steps ranging from zero to nine in value. This means that there is more than one sample to a scale-step. The first three and the last two steps have only one sample each, but all the intermediate ones have two or three each. The step intervals proceed regularly by units of one. This is another general merit scale with no expository

tion of the relative merits and defects of the respective scale-steps. The teachers' handbook which accompanies this scale contains three practice series of ten compositions each. The scorer may thus test out his accuracy in rating compositions by comparing his scores with those assigned by competent judges and reported in the score key to these practice exercises.

The latest scale published is that by S. A. Leonard, entitled, "A Scale for the Judgment of Composition Quality Only." This consists of ten scale-steps whose values are designated as scale-step 1 to 8.5, inclusive. The step intervals from 1 through 6 proceed by units of 1, and from thence forward by units of .5 more or less regularly. Almost every step contains more than one sample. Each sample is followed by a statement of its merits, defects, median, and inter-quartile range. This analysis justifies the position of each sample in the scale. The author admonishes the rater to be sure to overlook every matter of form as far as possible, and to consider only "the ideas; their interest and value (to the child, not to a school master); the effectiveness of planning the story to bring out the difficulty and worthwhileness; and the clearness and vigor of the sentences and the words used." In the words of Leonard,¹³ ". . . this scale is intended for use with the commoner sort of themes which do not correlate so conveniently (in thought and form), which are careful in form, but weak in ideas, organization, or other elements of pure composition quality, or which are excellent in these but deficient in formal matters." It essays to measure general merit aside from matters of form, and is applicable in grades IV through XII.

For a more detailed explanation of any further characteristics of the scales, of their derivation and the like, the writer refers the reader to the bibliography of this study.

In summarizing the nature of these scales perhaps the clearest picture can be gained from the sort of bird's-eye view afforded by Klapper's¹⁴ chart:

¹³ Leonard, S. A. *English Journal*, Vol. 14, December, 1925, p. 765.

¹⁴ Klapper, Paul. "Teaching English in Elementary and Junior High Schools." D. Appleton & Company, 1925, p. 220.

CLASSIFICATION OF COMPOSITION SCALES

Basis of Classification	The Scales
Those that classify according to general merit.	Hillegas, Thorndike, Nassau, Harvard-Newton, Breed and Frostic, Lewis, Hudelson.
Those that diagnose and analyze.	Willing, Van Wagenen.
Those that measure all types of writing.	Hillegas, Thorndike, Nassau, Breed and Frostic, Willing, Hudelson.
Those that measure special types of writing, like narration, exposition, letters.	Harvard-Newton, Van Wagenen, Lewis.
Those designed for use regardless of writing, like narration, exposition, letters.	Hillegas, Thorndike, Nassau, Willing, Lewis, Van Wagenen, Hudelson.
Those that apply to specific school grades.	Harvard-Newton (8th Grade), Breed and Frostic (6th Grade).
Specially adapted for junior high schools.	Nassau, Breed and Frostic, Willing, Van Wagenen, Hudelson, Lewis Narration.

COLLECTION OF COMPOSITIONS

Selection of composition topic. During the month of October, 1927, the Baltimore and Ohio Railroad celebrated the hundredth anniversary of its birth. To commemorate this event the company presented a pageant entitled "The Fair of the Iron Horse." It was an open air exhibition of the evolution of transportation in our country from the time of the rude ox cart to the luxury of the modern "Capitol Limited." Every school in the city spent a day viewing the pageant and examining the various exhibits on the exposition grounds. Since there was very general enthusiasm and knowledge concerning the subject in the whole City of Baltimore, the writer deemed this topic the best possible available one to provoke the school children's hearty co-operation and interest in furnishing compositions for this experiment. This general subject was divided into five aspects, entitled as follows:

1. The Funny Things I Saw at the Fair of the Iron Horse.
2. The Best Way to Get to the Fair of the Iron Horse.
3. A letter to a friend, telling him or her what a good time you had at the Fair of the Iron Horse.
4. A letter to a friend who did not wish to go to the Fair of the Iron Horse. Tell how you got him or her to go.
5. A letter to the director of the pageant, asking for a ticket for a reserved seat in the grand-stand.

The children were asked to choose only one of these subjects as the topic for their composition. Conditions of writing these and the instructions given the pupils by the experimenter will be discussed later. The reason for giving this variety of themes was to supply each type of discourse and special kinds of writing (such as letters) to fit the various composition scales upon which they were to be marked. Thus, topic (1) was calculated to evoke a description; topic (2) to bring forth an exposition; topic (3) to produce a narration and, at the same time, a "narrative social" letter; topic (4) an argumentation and, at the same time, an "expository social" letter; and (5) an "order letter." By this arrangement compositions employing each of the four types of discourse became available for application of such scales as the Harvard-Newton and the Van Wagenen; compositions exemplifying three types of letter writing were available for application of the Lewis Scale. It may be urged that letters asking for a ticket are a poor sort of order letter. However, unsatisfactory as this is, it was the only approach to that type of letter afforded by the general topic: "The Fair of the Iron Horse."

Subjects of the experiment. Children in the 6B grade (low sixth) in six different schools were asked to write these compositions under the personal direction of the experimenter and without advance knowledge of the test. The schools were chosen on the basis of the various strata of society which they typified. With one exception they were all large schools. The first was representative of the semi-professional and middle classes; the second, of the non-professional middle classes; the third, of the professional and "better-to-do" classes; the fourth, of the middle and poorer classes; the fifth, of the poorer industrial element of the city; and the sixth, of the suburban and wealthier classes.

Conditions of testing. Standard conditions obtained throughout the administration of the composition writing. Every school was visited during the week of October twentieth. All participating classes (eleven in number) wrote compositions during the first two hours of the morning ses-

sion with the exception of two. These wrote the first period after lunch. The experimenter entered the room and asked the teacher to see that each child was provided with ruled paper and a well sharpened pencil or pen. Further, each child was to draw a half-inch margin on the left-hand side of each page of the paper, and then to sit still until the experimenter was ready. In the meantime the experimenter had written on the board the main title, followed by the five sub-titles as herein previously stated.

Then the pupils were asked to give the following information, one item at a time:

(a) Name—(b) Address—(c) Age—(d) Number of school and date—(e) Name of teacher—(f) Grade and room number—(g) Have you seen the Fair of the Iron Horse? If you have seen the entire pageant from beginning to end write "yes." If it rained when you went, and you saw only a part of it, write "part."

Now turn your paper over:

(h) I want you to write me a little composition about the Fair of the Iron Horse. You will be given exactly 15 minutes in which to write the best story you can on this subject. (i) Do not begin until I say "ready, begin." (j) On the top line write your title, "The Fair of the Iron Horse." (k) Lay down your pen and listen. (l) Now I am going to give you some ideas as to what you might tell me about this pageant, and you may choose one of these suggestions. In your composition tell about *ONE* of the following:¹⁵ (m) I shall give you a moment to pick one out. (Waited about one minute.) Remember, write on only one of these subjects. Now write the number of the one you have chosen beside your title. (Showed them on board.) (n) Now, ready, begin. (Looked at watch and recorded time.) At the end of fourteen minutes the experimenter reminded the pupils that they had one more minute to finish up, and that they were to stop promptly at the end of the minute. After fifteen minutes the experimenter said, "Everybody stop," and the teacher took charge of collecting the papers.

In addition to the above information which the experimenter required each child to give on his paper, the former procured the I. Q., the XYZ grouping, the reading rate, the reading comprehension score and the rank of each pupil in the class by consulting the class analysis chart.

Present-day use of composition scales. In order to determine the extent to which educators all over the country actually use these composition scales, the following questionnaire

¹⁵ See sub-titles previously enumerated in this chapter under the heading: Collection of Compositions.

was sent to representative educational authorities over the whole United States.

USE OF COMPOSITION SCALES

Name—Teaching position—What grade do you teach?—What subjects do you teach?—In what school do you teach?—In what city?—County?—State?

1. Have you ever used a composition scale in the marking of the compositions of your class?

a. In which grade or grades?

NOTE.—If you have never used a scale for grading compositions, please answer questions 1 and 7 nevertheless, and send in your questionnaire. It will be helpful.

2. Below is a list of available scales. Check in the right margin those you used.

a. Nassau County Supplement to the Hillegas Scale, by Trabue.

b. Harvard-Newton Composition, by Ballou.

c. Hudelson English Composition Scale. (Sometimes known as "The Virginia Supplement to the Hillegas Scale"), by Hudelson.

d. Typical Composition Ability Scale, by Hudelson.

e. Maximal Composition Ability Scale, by Hudelson.

f. Minnesota English Composition Scales, by Van Wagenen. (Sometimes known as the Van Wagenen English Scale.)

g. Willing's Scale for Measuring Written Composition, by Willing.

h. Scales for Measuring Special Types of English Composition, by Lewis.

i. Scale for Measuring Sixth Grade Composition, by Breed and Frostic.

j. Thorndike Extension to the Hillegas Scale.

k. Hillegas Scale for the Measurement of Quality in English Composition for Young People.

1. Leonard's Scale for the Judgment of Composition Quality Only.

3. Which, in your judgment, is the best scale?

4. On which of the following reasons did you base your judgment?

a. It takes the least time.

b. It is the least complicated to use.

c. It helps most in bringing out the weak points in pupils' compositions.

d. It conforms most nearly with the teacher's own judgment.

e. It is a truer measure of the real merit of the composition.

f. List below any other reasons for your scale preference, mentioning the scale.

5. How often do you mark compositions by a scale?

6. When using the scale, do you have the pupils write on the same topic or topics as the samples which appear in the scale?

a. Do you apply the scale to any topic?

7. Do you consider composition scales of any value?

a. Why?

To further corroborate the findings of the writer's own questionnaire, seventy-eight surveys made between the years 1912, the date of publication of the very first composition scale, and 1928, inclusive, were examined for their employment of composition scales.

CHAPTER II

TECHNIQUE OF THE EXPERIMENT

The main steps in the technique of this experiment are :

1. The marking of one hundred compositions on each of seven composition scales by twenty-five judges untrained in the use of such scales.
2. A period of three to three and one-half hours' training on standardized practice material with periodic comparison of judgment to standard values.
3. A repetition of "1" above after training.
4. Subjective ranking of scales in order of merit by each judge.

SELECTION OF COMPOSITIONS

A total of three hundred and fifty compositions was written by the eleven 6B classes participating in the experiment. It was decided to choose a random sample of one hundred from among these. The choice was made upon the basis of chronological age, intelligence quotient, XYZ grouping, type of composition topic, and class roll. In order to do this, a sheet containing the pupil's name, rank in class assigned by the teacher, chronological age, I. Q., XYZ group, reading rate, reading comprehension and number¹⁸ of composition was made out. The caption of these sheets consisted of the teacher's name, grade, school, date of reading scores and of composition writing. Each of these sheets summarized the number of XYZ groups, their corresponding intelligence limits, and the grade levels in reading scores. These eleven sheets were alphabetically arranged and were the basis for most of the future classification. The age distribution of each teacher's class was made for the total of three hundred and fifty pupils.

A sheet giving the intelligence distribution of all classes was also made for the entire group. A surface frequency distribution of the chronological ages of the three hundred

¹⁸ See Chapter I, footnote 15.

and fifty 6B pupils as of September, 1927, was made. The total number of 6B children in the grades tested was given, the number of those present who wrote compositions, and the number of those absent who wrote no compositions. These findings are reported in Table I.

A distribution of composition topics, class by class, was made. (See Table I.)

TABLE 1
DISTRIBUTION OF COMPOSITION TOPICS

Teacher	Class roll	No. of comps.	Composition topic numbers				
			1	2	3	4	5
A	47	40	19	8	9	3	1
B	17	17	11	3	2	0	1
C	19	16	8	4	4	0	0
D	28	28	15	7	6	0	0
E	46	43	14	9	14	3	3
F	33	32	17	8	3	3	1
G	45	39	18	10	10	1	0
H	34	34	17	8	6	1	2
I	42	36	17	9	9	0	1
J	29	26	12	5	7	1	1
K	41	39	12	7	14	3	3
Totals ²⁷	381	350	160	78	84	15	13

Since only thirteen pupils out of a total of three hundred and fifty chose composition topic number 5, and only fifteen selected topic number 4 as their subjects, it was decided that all of these would be included in the final hundred. A deduction of these numbers from one hundred left seventy-two compositions to be equally distributed among the other three composition topics. This made an assignment of twenty-four compositions for each of topics 1, 2, and 3.

For the purpose of securing a really random sample, the writer consulted the Baltimore Bureau of Research for the "Age-Grade Progress Table" for 6B as of February, 1927, for white children in Baltimore. The number of boys and girls at each grade was added, and the per cent which each age represented of the total was determined. The propor-

²⁷ The difference between the total roll and the total number of compositions written is due to the fact that thirty-one pupils were absent, and, hence, wrote no compositions.

tions in which our hundred pupils would have to be chosen in order to represent the proper per cent of each age group were calculated, and are given in Table 2.

TABLE 2
AGE-GRADE PROGRESS TABLE FOR 6B WHITE PUPILS IN BALTIMORE
AS OF FEBRUARY, 1927

Age	Boys	Girls	Both	Per cent	No. of pupils
9 yrs.	0	1	1	.02	1
9½	7	4	11	.21	1
10	44	40	84	1.65	1
10½	113	123	236	4.64	4
11	332	379	711	13.99	14
11½	467	554	1021	20.10	20
12	482	473	955	18.80	18
12½	354	318	672	13.23	13
13	286	249	535	10.53	10
13½	220	173	393	7.73	7
14	142	86	228	4.45	4
14½	88	48	136	2.67	3
15	35	18	53	1.04	1
15½	19	11	30	.59	1
16	7	4	11	.22	1
16½	1	1	2	.04	1
17	0	0	0	.00	
Total	2597	2482	5079	99.91	100

In order to facilitate the selection of one hundred compositions from the total of three hundred and fifty, the class sheets were gone through, and a sheet giving all the "X-plus pupils" in all of the eleven grades with a notation of their chronological age, I. Q., and composition topic was made. The same was done for X, Y-plus, Y, Y-minus, Z, and Z-minus pupils.¹⁸ The groupings are based upon definite limits of I. Q. These limits,¹⁸ together with the per cent of pupils chosen from each, are:

I. Q.	Group	Per cent of pupils chosen
120-158	X-plus	8
110-118	X	12
102-108	Y-plus	18
88-100	Y	24
78-86	Y-minus	18
68-76	Z	12
60-66	Z-minus	8
and under		

¹⁸ Taken from class analysis charts of Baltimore City Schools.

In order to get a purely random sample with respect to chronological age, intelligence and class roll (as far as possible), the following procedure was adopted:

1. Included all compositions on topic number 5.
2. Prepared a sheet containing the names of the pupils, the teacher, chronological age, and I. Q. group. Summed up the distribution of XYZ and chronological age groups. Included all of number 4 compositions.
3. Subtracted the actual number of pupils of each chronological age in each group from the total number desired. Apportioned these (both the intelligence and chronological age groups) equally among the other three composition topics, trying to get the same number of each group and the same number of each chronological age for each composition topic. Similar sheets prepared for topics 4 and 5 were made for composition topics 1, 2, and 3. The results of the total distribution for each composition topic, age, and intelligence was made on one sheet, and the actual number compared with the desired number of each of the groups in question.

Table 3 gives the actual number of children at each age writing the respective composition topics.

TABLE 3

AGE DISTRIBUTION WITH RESPECT TO COMPOSITION TOPIC CHOICE

Comp. topic	Chronological Ages																Total
	8½	9	9½	10	10½	11	11½	12	12½	13	13½	14	14½	15	15½	16	
1	0	0	0	0	0	3	3	4	4	3	5	1	0	1	0	0	24
2	0	0	0	0	0	3	4	5	3	2	2	2	1	1	1	0	24
3	0	1	0	0	1	2	4	6	4	2	1	1	1	1	0	0	24
4	0	0	1	2	2	3	3	4	0	0	0	0	0	0	0	0	15
5	0	0	0	1	1	4	5	0	1	1	0	0	0	0	0	0	13
Actual total	0	1	1	3	4	15	19	19	12	8	8	4	2	3	1	0	100
Desired total	0	1	1	1	4	14	20	19	13	10	8	4	3	1	1	0	100

Table 3 should be read as follows:

One 9-year-old wrote on composition topic number 3. One 9½-year-old wrote on composition topic number 4, and so on.

Table 4 gives the actual number of children at each intelligence level writing on the respective theme topics. This should be read: One X-plus pupil wrote on composition topic number 3. Six X-plus pupils wrote on composition topic number 4. Three X-plus pupils wrote on composition topic 5, and so on.

TABLE 4

INTELLIGENCE DISTRIBUTION WITH RESPECT TO COMPOSITION TOPIC CHOICE

Comp. topic	XYZ groups							Total
	X-plus	X	Y-plus	Y	Y—	Z	Z—	
1	0	3	3	6	5	7	0	24
2	0	2	4	11	6	1	0	24
3	1	2	4	8	6	2	1	24
4	6	3	2	4	0	0	0	15
5	3	2	5	2	1	0	0	13
Actual total	10	12	18	31	18	10	1	100
Desired total	8	12	18	24	18	12	8	100

The above procedure insured a random sample of chronological ages, intelligence groups, and composition topics in proportions approximating rather closely the normal curve of distribution.

PREPARATION OF COMPOSITIONS

When the hundred had been chosen as described, type-written stencils were made, retaining the original spacing and errors of all kinds. It was thus attempted to reproduce everything in the original composition except the handwriting. These stencils were mimeographed and cut into uniform sheets containing one composition each. These were put into packages of one hundred and fastened. In order to prevent recognition of a composition by a judge as he marked it on each scale, before and after training, the judges were instructed to put no score or identification marks of any sort on the composition. Further, a code number of the kind which could not be remembered by the judge was assigned to each composition. Instead of numbering the compositions from 1 to 100, these numbers were given alphabetical equivalents. Z represented 1; Y, 2; X, 3; W, 4; V, 5; U, 6; T, 7; S, 8; R, 9; and Q, 0. To further prevent recognition, Q (which equals 0) was prefixed before every code number except 100, in order that each code might have three letters

in it. For a complete identification of the composition the experimenter made each complete code number consist of four groups of letters. The first of these was the number of the composition; the second represented the child's initials; the third, the teacher's initials; and the fourth, the topic which the composition exemplified. Thus, the complete code number of the first composition was *QQZ-AA-MG-d*.

The final letter at the end of each code number represents the composition topic chosen, as follows: *a* stands for composition topic number 1; *b* for 2; *c* for 3; *d* for 4; and *e* for 5.

SELECTION OF JUDGES

The experiment required twenty-five people who were entirely unfamiliar with composition scales to act as scorers or judges of these compositions. Most of the judges were secondary school (Junior High School) English teachers, graduate students in the Education Department of The Johns Hopkins University, and under-graduates in the same school, as well as in the College of Liberal Arts of the same institution. It was found that most of the under-graduate students participating were planning to be teachers of English. Several of these had been trained for elementary school work in the Normal School at Towson, Maryland, and had experience in teaching the intermediate grades. One supervisor of these grades also took part in the experiment. Both sexes were represented among the judges.

DIRECTIONS TO JUDGES

Each judge was given a manilla envelope containing a mimeographed sheet of general directions, a sheet of punctuation rules, a copy of the particular scale to be used, a set of compositions, and a score sheet. A sample of each of these, with the exception of the compositions, is contained in the appendix. (The sheet of punctuation rules was included in order that the same grammar rules might be followed by all in marking "form errors" on the Willing Scale.) The judges were instructed to record on the score sheet the time

it took them to become familiar with the particular scale in hand, the number of compositions they marked each hour, and the total time spent in marking one hundred of them. Room was also provided on this sheet for the judge's personal re-action and criticism of the particular scale after using it. In addition to this, the score sheet contained the complete code numbers of the hundred compositions in the particular order in which they happened to be arranged. This order was changed each time the judge received a new scale to prevent memory of the score assigned on the previous scale. Opposite each code number on the score sheet, the judge was instructed to record the score assigned to the given composition. In rating letters on all scales except the Lewis, judges were directed to consider only the body of the letter, and rate this as they would any other composition. Thus, the letter form was rated on the Lewis only, while the body of each theme was rated on each of the seven scales. The package was then to be returned to the experimenter, the compositions were shuffled, and a new score sheet and scale were inserted. This was repeated for each of the seven scales both before and after training.

A slip of paper was attached to each composition scale, calling the attention of the judge to the printed instructions which the author of each scale expects scorers to follow, and giving further specific directions adapting their use to the particular compositions in hand. These directions for each scale are given in full in the appendix. Only three specific facts need to be mentioned at this point. Wherever a scale made discourse distinctions the judges were to mark each set of composition topics on the appropriate scale. Since our total of one hundred compositions contained only fifty-two letters, the judges were instructed to apply the Lewis Scale to these only; and since Van Wagenen supplies no argumentation scale, the judges were told to omit the fifteen compositions representing this type of discourse, and, thus, to apply this scale to only eighty-five compositions.

PRACTICE EFFECT

Practice effect was eliminated by means of rotating the scales and enforcing this system of rotation by requiring the judges to finish marking the compositions on one scale before they were given the next. By this method each scale occupied the same position an equal number of times, as nearly as it was possible to arrange this. In other words, judge A, for example, had, as his first scale, the Nassau Supplement to the Hillegas Scale; judge B had the Willing for his first; judge C, the Hudelson Typical, and so on. Judge A had for his second scale the Willing; judge B had the Nassau; and judge C had the Lewis. Thus, the facility gained by inexperienced judges from using the seven scales in different order was approximately evenly distributed from scale to scale, and practice effect was thus ruled out.

After training, the succession in which the scales followed each other was again re-distributed so that each scale occupied the same position practically the same number of times, but followed each other in an order different from that before training.

TRAINING OF JUDGES

Since the most widely used practice material for training in composition scales was the "150 Specimens Arranged for Use in Psychological and Educational Experiments," arranged by Thorndike,¹⁹ this series was employed also in this experiment. Owing to the fact that sufficient copies were not available for all the judges, the latter were given mimeographed sheets, reproducing the compositions as given in Thorndike's book. Attached to this series of sheets was a copy of Thorndike's Extension of the Hillegas Scale, and a record sheet containing spaces for the accommodation of the scores assigned to the "training compositions." The following directions were given:

1. First of all, become thoroughly familiar with the *whole* of Thorndike's Extension of the Hillegas Scale.
2. Then, watching the time carefully, work for one hour WITHOUT INTERRUPTION.

¹⁹ Thorndike, E. L. "English Composition." Teachers College, Columbia University, New York City, 1916.

3. Record the marks you assign to the "Training Compositions" on the accompanying record sheet. You will notice that the "Training Compositions" are arranged by sets designated as A, B, C, D, E, F, G, H, I, and J, respectively. In each set there are ten compositions. Go right through these, assigning marks according to the Thorndike Scale. If you should finish the series before the hour is up, begin at the beginning, and repeat the process, BUT COVER OVER THE MARKS ASSIGNED THE FIRST TIME to prevent prejudice. Do only enough to finish out the hour.

4. Finally, return the package to the experimenter and she will tell you how you deviate from expert judgment—either too severe, or too lenient—and the average number of points of deviation for the hour's training.

5. Train yourself a second hour by marking these same compositions and TRYING TO CORRECT THIS INDIVIDUAL TENDENCY OF YOURS.

6. On this scale you may use interpolated marks if you think the intermediate values ought to be assigned.

As the directions indicate, the judges consulted with the experimenter at the end of their first hour of training, at which time they were told what fraction of the compositions were marked too leniently, what proportion too severely, and whether they tended toward leniency or severity as indicated by a comparison of the judges' scores to those in the scoring key. In each case the average deviation (*plus*, to indicate leniency, and *minus*, to indicate severity) was given to the judge. The next day the judge took another hour of training. At the end of the second hour of training, the judge again consulted with the experimenter, and his deviation, as described above, was again calculated and revealed to him.

The following day the judge took one-half hour of training. During this time the judge began marking compositions from the place where he had left off (rather than repeating, as he did the second hour), and marked as many as possible within thirty minutes. At the end of this time the judge again consulted with the experimenter and was told his tendency, as previously.

On the following day, a different set of practice material was used. It consisted of the first five samples of known value chosen from the "Samples for Practice in Rating" provided by Hudelson in his English Composition Scale, page 6; the first five provided by Hudelson in the Teachers' Handbook accompanying "The Hudelson Typical Composi-

tion Ability Scale," page 9; and the first five of "Selected Specimens for Practice in Rating Themes" provided by Van Wagenen in the "Van Wagenen English Composition Scales" for each of the three discourses (exposition, narration, and description) pages 13, 16, and 18, respectively. This made a total of twenty-five compositions which were to measure the "effect of training." A sheet entitled "Test of the Effect of Training" was given to each judge, and the following directions were stated:

1. Record marks in the proper space below for the first five samples in the practice series of each of the three scales mentioned.
2. Interpolate scores on all three scales.
3. Do not work out the general merit on the Van Wagenen Scale.
4. Do the samples in the order indicated.
5. While marking these practice samples keep in mind your general tendency, revealed by your two and one-half hours' practice, and try to correct this.
6. Record the total time it took you to mark the twenty-five compositions.

Here, as was done throughout the whole experiment, the three scales (Hudelson English, Hudelson Typical, and Van Wagenen) were rotated so that approximately the same number of judges had these scales in each of the three positions the same number of times. This is what was cared for in direction number 4.

The reason for choosing these three scales for testing out the transfer of improvement resulting from two and one-half hours' practice on the Thorndike material was: first, the fact that of all the seven scales tested out, these three were the only ones provided with practice samples within their folders; and, second, the fact that the two Hudelson scales are Hillegas derivatives, while the Van Wagenen is not. Hence, it could be argued that the former could be expected to improve more as a result of training on a related scale such as the Thorndike, while the latter would not improve to the same degree.

When the judge finished this "Test of the Effect of Training," he consulted, for the last time, with the experimenter, and was told, as previously, his average deviation from expert judgment, and his tendency toward severity or leniency,

as the case might be. He was admonished to bear this tendency in mind when repeating marking the hundred compositions on the seven composition scales after training. In most cases this test required from one and one-half to two hours, thus making the total time spent in training four hours. This amount of time was all that was available.

After this period of training the twenty-five judges followed the same procedure as they had followed before training. Any difference in results, therefore, is due to the effect of training.

SUBJECTIVE RANKING OF SCALES BY JUDGES IN ORDER OF PREFERENCE

When all ratings of the compositions both before and after training were completed, each judge was given a questionnaire in which he was required to rank the seven composition scales assigning number one to the scale which he or she considered best, two to the next best, and so on. After this, the list of possible reasons for merit or short-comings in composition scales was suggested, and the judge was instructed to check those which made him consider the one marked 1, the best, and to give a separate list of those which made him consider the one he marked 7, the worst. At the end of the questionnaire, room was provided for the judge to register any other reasons controlling his preferences not covered in the list of suggested ones. This was intended to supplement the individual criticisms of each scale which each judge gave on each score sheet (as described in the "Directions to Judges" earlier in this chapter).

CHAPTER III

CRITERIA USED IN THIS STUDY

Three external criteria of objectivity are used in this study. The first of these is the size of the coefficient of variation of each scale. Improvement resulting from training is measured by the shrinkage in the coefficient of variation; in other words, the difference between the coefficient of variation of a given scale before training and its corresponding value after training. This is designated as Criterion I, and was derived by taking the average of one hundred coefficients of variation (one for each of the hundred compositions) for each of seven scales. Results arrived at by application of this criterion will be given in Chapter IV.

The second criterion employed was the rank coefficient of correlation known as *rho*. This was built up by selecting from the hundred compositions forty whose quality gives an approximately normal curve of distribution on the Hudelson English Scale after training.²⁰ This curve is given in Chapter IV. These forty compositions were selected according to their quality on this same scale, and were ranked in order of merit, assigning rank 1 to the theme of highest quality. Assuming that the quality of a composition is fairly constant²¹ irrespective of the scale by which it is measured, each of these specimens was ranked according to its quality on all of the seven scales except the Lewis. The reason for omitting the last named is the fact that this scale is one for letters only. There were just fifty-two such letters included in our total of a hundred. Fifteen of these represented the

²⁰ The Hudelson English Scale served as the basis of selection because, according to Criterion I, it is the most objective of the scales investigated after training.

²¹ For justification of this assumption see Chapter IV.

argumentative type of discourse for which another scale, namely the Van Wagenen, provided no measurement. This left only thirty-seven compositions which were marked by each one of the seven composition scales, and it is hardly probable that these would have distributed themselves normally. Hence, for the sake of random sampling, and a sufficiently large number to insure reasonable reliability, forty compositions were chosen and marked on each scale except the Lewis.

The sum of the ranks which these forty compositions took on each of the six scales was computed. These sums were again re-ranked in order of merit, giving the lowest sum rank 1. This represented the criterion rank. Then *rho* was found for each scale by taking the difference between the rank of each composition on a given scale and its corresponding criterion rank, squaring this, and applying the formula

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

Results will be given in the following chapter.

The third criterion used was "chance."²² Most authorities²³ agree that composition scales reduce variability of judgment, especially where such judgment is rendered by persons trained in the use of composition scales. Most of these point out the advantage of the scores assigned by scales over those assigned on the commonly used per cent system, or speak of the relative superiority in objectivity. So far as the writer is aware, however, no one has shown the difference in variability of the composition scales in absolute units. The greatest possible variability in judgment likely to be obtained would be that resulting from assignment of values by chance. This has been calculated for each scale and is given in Table

²² Brooks, F. D. "Relative Accuracy of Ratings Assigned With and Without Use of Drawing Scales." *School and Society*, Volume 27, April 28, 1928, pp. 518-520.

²³ See articles by Van Wagenen, Theisen, Briggs, and Thorndike given in the bibliography.

II. The standard deviation of chance ratings is then compared with the average standard deviation of each of the seven scales. Thus, the decrease in variability of ratings assigned by the use of each of the scales from its corresponding chance variability is expressed in absolute units, for it will thus be possible to state *how much* more objective a composition scale is than the worst possible standard. This chance criterion is designated as Criterion III.

CHAPTER IV

RESULTS OF THE EXPERIMENT

The marks assigned to each of the hundred compositions by twenty-five judges were transcribed from the score sheets to a work chart (one for each scale before training, and a corresponding one after training). These charts contain a total of fifty-three thousand marks. For each composition its average value (composite value assigned by twenty-five judges), standard deviation, and the Pearson coefficient of variation were calculated. The last named seeks to eliminate the differences of units in which the various scales are measured. This was done for each of the seven scales, both before and after training. These data, in condensed form, are found in Table 5. In this table the hundred compositions were grouped into five groups: the first three containing twenty-four samples each; the fourth, fifteen; and the fifth, thirteen. The groups are arranged in the order of the composition topics on which the pupils wrote as explained in Chapter I. The arithmetic average of the means, standard deviations, and coefficients of variation were calculated, and it is these which are given in Table 5. Use will be made of these data later.

OBJECTIVITY OF SCALES AS DETERMINED BY CRITERION I

The coefficients of variation of the hundred compositions were averaged for each scale before and after training, and the improvement as a result of training was indicated by the difference between the two. Results are given in Table 6. All conclusions must be interpreted within the limitations of this criterion, which is the extent to which the Pearson coef-

TABLE 5²⁴

MEAN, STANDARD DEVIATION, COEFFICIENT OF VARIATION OF 100
COMPOSITIONS GROUPED BY TYPE OF DISCOURSE FOR
EACH SCALE

	No. of Comps.	SCALE					
		Hudelson English			Hudelson Typical		
B.T. ²⁵		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des. ²⁵	24	3.14	.72	22.98	2.69	.99	38.02
Exp. ²⁵	24	2.96	.70	23.80	2.48	.91	37.27
Nar. ²⁵	24	3.30	.75	22.94	2.79	.94	34.58
Arg. ²⁵	15	3.41	.80	23.59	2.97	1.01	34.77
O. L. ²⁵	13	3.56	1.01	28.13	3.26	1.13	35.04
		Leonard			Lewis (Letters)		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des.	24	3.63	.96	27.21			
Exp.	24	3.10	.86	28.61			
Nar.	24	3.56	.93	27.51	33.05	9.02	28.40
Arg.	15	3.61	.91	25.86	33.14	10.14	31.07
O. L.	13	3.30	.93	28.61	35.23	11.68	33.36
		Nassau Supplement			Van Wagenen		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des.	24	3.35	.86	26.06	29.64	8.44	29.30
Exp.	24	3.02	.83	27.58	25.53	7.59	30.20
Nar.	24	3.39	.84	25.55	29.55	8.48	28.93
Arg.	15	3.63	.91	25.15			
O. L.	13	3.73	1.03	27.71	33.91	13.09	38.68
		Willing—Story Value			Willing—Form Value		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des.	24	36.06	9.36	26.11	13.97	3.89	29.57
Exp.	24	33.90	10.03	30.85	13.01	4.07	32.83
Nar.	24	36.78	9.60	26.30	13.71	4.14	32.66
Arg.	15	36.66	11.06	30.39	11.72	4.24	37.87
O. L.	13	35.75	14.91	41.59	10.12	4.39	46.72
		Hudelson English			Hudelson Typical		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
A.T. ²⁶							
Des.	24	3.56	.70	19.97	3.20	.83	26.44
Exp.	24	3.36	.75	22.30	2.96	.85	28.80
Nar.	24	3.68	.73	19.97	3.25	.84	26.24
Arg.	15	3.82	.82	22.03	3.39	.87	26.22
O. L.	13	4.02	.87	21.77	3.61	.99	27.42
		Leonard			Lewis (Letters)		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des.	24	4.26	.87	20.96			
Exp.	24	3.75	.87	23.55			
Nar.	24	4.24	.86	20.80	39.10	9.39	25.13
Arg.	15	4.25	.92	22.38	36.41	7.78	22.07
O. L.	13	4.09	1.26	31.00	39.61	12.42	31.72
		Nassau Supplement			Van Wagenen		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des.	24	4.03	.95	24.10	36.06	8.06	23.07
Exp.	24	3.74	.95	25.74	30.48	8.21	27.34
Nar.	24	4.05	.95	23.71	34.25	7.96	23.84
Arg.	15	4.21	.98	23.48			
O. L.	13	4.33	1.09	25.40	37.98	12.57	33.21
		Willing—Story Value			Willing—Form Value		
		Mean	S. D.	C. V.	Mean	S. D.	C. V.
Des.	24	41.38	10.56	26.29	13.21	3.08	31.92
Exp.	24	36.63	10.64	29.23	11.87	3.85	33.68
Nar.	24	41.83	10.32	25.38	12.71	4.26	35.87
Arg.	15	41.30	11.23	27.99	11.13	4.32	40.88
O. L.	13	41.38	14.06	34.02	9.41	4.82	52.22

²⁴ Complete tables giving these data for each of the hundred compositions are to be found in the manuscript copy on file in the library of The Johns Hopkins University.

²⁵ Explanation of abbreviations:

B. T.—Before Training. Des.—Description. Exp.—Exposition. Nar.—Narration. Arg.—Argumentation. O. L.—Order Letters. C. V.—Pearson Coefficient of Variation.

²⁶ A. T.—After Training.

ficient of variation really does eliminate the difference in size of units of the various scales.

TABLE 6

AVERAGE COEFFICIENT OF VARIATION OF EACH OF SEVEN COMPOSITION SCALES

Scale	Before training	After training	Gain
Hudelson English	23.93	21.07	2.86
Hudelson Typical	36.14	27.05	9.09
Leonard ²⁷	27.60	23.06	4.54
Lewis	30.41	25.89	4.52
Nassau	26.38	24.48	1.90
Van Wagenen	30.88	26.04	4.84
Willing ²⁸	29.95	28.04	1.91

The reader will recall that the size of the coefficient of variation was taken as Criterion I—the smaller the coefficient of variation, the better the scale. According to this, the scales before training rank as follows in their objectivity:

1. Hudelson English, 2. Nassau Supplement, 3. Leonard, 4. Willing, 5. Lewis, 6. Van Wagenen, 7. Hudelson Typical.

According to this ranking of the scales the Hudelson English is the most objective, the Nassau Supplement the next most objective, and so on.

After training, according to the same criterion, these scales take the following ranks:

1. Hudelson English, 2. Leonard, 3. Nassau Supplement, 4. Lewis, 5. Van Wagenen, 6. Hudelson Typical, 7. Willing.

It would seem from this that trained judges will probably assign the most objective marks on the Hudelson English, and the least on the Willing.

²⁷ Before training it was possible to get only nine judgments on each of the hundred compositions because the scale was not included in the study until it had proceeded for some time. After training, however, each composition received 25 judgments.

²⁸ Besides the mark for story value, each composition was given a mark for form value. According to the directions of Willing this is the number of mistakes per hundred words. For purposes of comparison only Willing "Story Value" is considered. (See Denver Survey, p. 61.)

It is clear that training decreases the variability,²⁹ and, hence, increases the objectivity of every scale.

The same results are demonstrated perhaps more clearly by the curves in Figures 1 to 8, inclusive. Each of these graphs consists of one hundred coefficients of variation (one for each composition) for each scale before, and the same after, training. The arithmetic mean and the standard deviation for each of these graphs is given in Table 7.

TABLE 7

MEAN COEFFICIENTS OF VARIATION AND STANDARD DEVIATIONS
CORRESPONDING TO FIGURES 1-8

Scale	Before training		After training		Gain
	C. V.	S. D.	C. V.	S. D.	
Hud. English	23.93	4.31	21.07	3.32	.99
Hud. Typical	36.14	7.19	27.05	4.90	2.29
Leonard	27.60	7.24	23.06	5.89	1.35
Lewis	30.41	6.17	25.89	7.45	— 1.28
Nassau	26.38	4.57	24.48	4.04	.53
Van Wagenen	30.88	6.39	26.04	6.23	.16
Willing ³⁰	29.95	7.53	28.04	5.61	1.92

The reader should note that these standard deviations are directly comparable from scale to scale because they are

²⁹ It is interesting to note that, according to Criterion I, the Hudelson English Composition Scale and the Nassau County Supplement to the Hillegas Scale are the most objective scales of those examined before training, and, with a single exception, after training. In view of this, untrained as well as trained scorers will assign more objective marks on these than on the others. The reader should note that the decrease in the size of the coefficient of variation as a result of training is less in these (except the Willing) than in the other scales. This, together with the fact that their coefficients of variation are among the smallest, would indicate that, wherever circumstances preclude the opportunity for training, scorers will do well to choose one of these two scales. While the Willing Composition Scale has one of the smallest of the coefficients of variation before training, and makes little decrease as a result of training, the fact that its variability is among the highest after training makes it one of the least objective scales. What has been said applies to the Willing "Story Value" only. The "Form Value" yields less objective results than any of the above scales: C. V. before training being 34.57, after training 37.27, thus yielding a negative gain; namely, — 2.70. This seems to show that the Willing Form is not at all objective. The standard deviation before training is 11.18, and after training 11.54. Figure 8 shows the same results.

³⁰ See footnote number 29.

measured in similar units; namely, the coefficients of variation. An inspection of the variabilities of these scales, as measured by their standard deviations, indicates unmistakably that trained judges are more consistent in their ratings than untrained ones. It also indicates that untrained markers can use both the Hudelson English and the Nassau scales to better advantage than the others; similarly, that the Hudelson Typical and the Leonard require training if they are expected to yield objective results. This is in agreement with conclusions previously drawn.

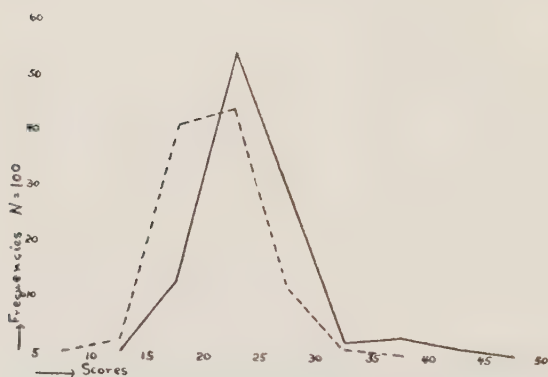


Figure 1 - Hudelson English Coefficients of Variation

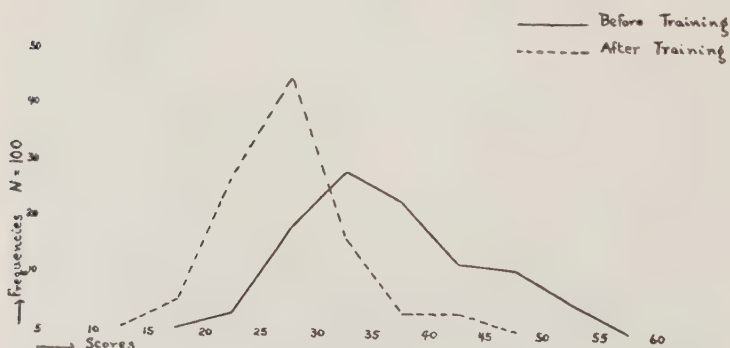


Figure 2 - Hudelson Typical Coefficients of Variation

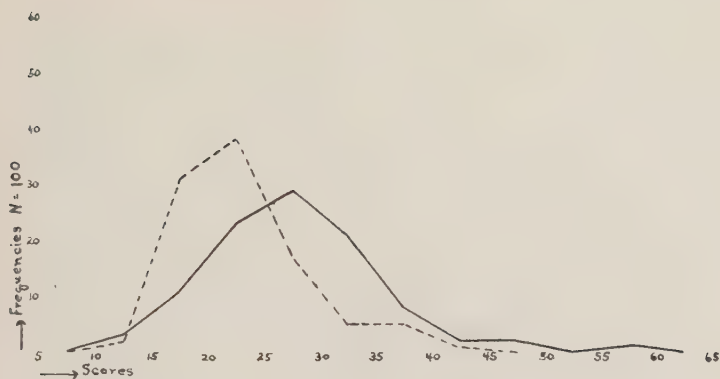


Figure 3 - Leonard Coefficients of Variation

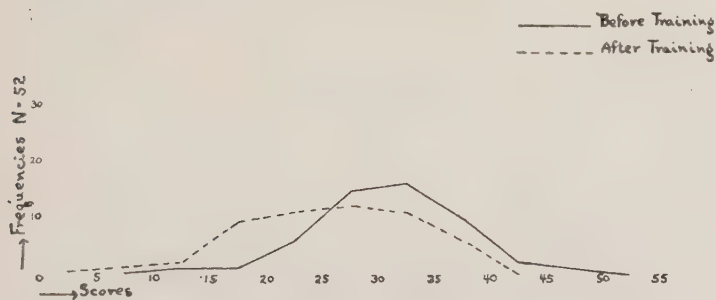


Figure 4 - Lewis Coefficients of Variation

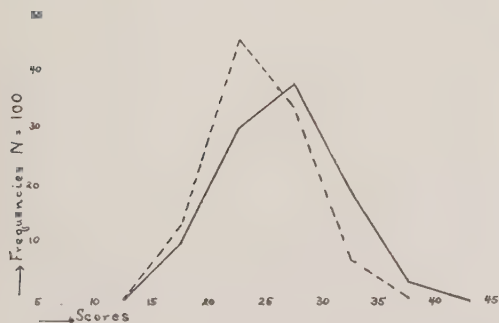


Figure 5 - Nassau Coefficients of Variation

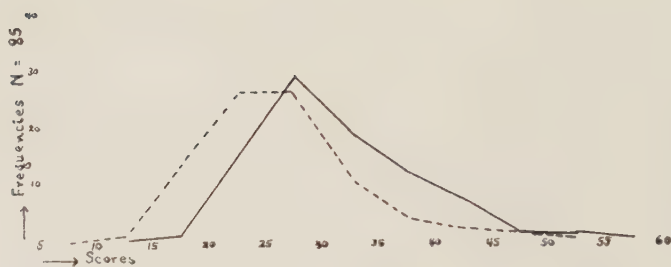


Figure 6 - Van Wagenen Coefficients of Variation

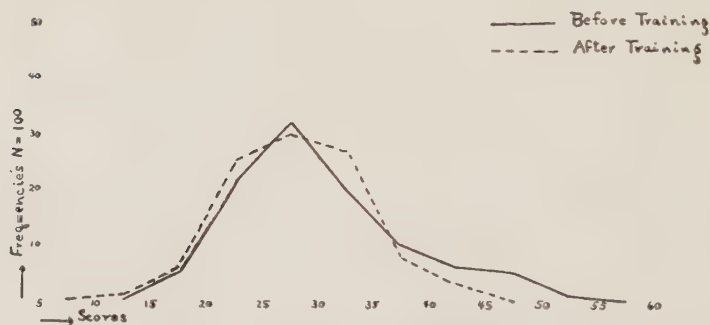


Figure 7 - Willing - Story Value - Coefficients of Variation

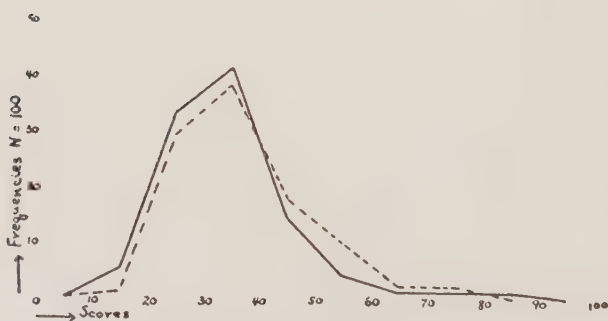


Figure 8 - Willing - Form Value - Coefficients of Variation

In order to discover what improvement or gain in objectivity can result in compositions of different values, fifty specimens were selected from the total of one hundred for intensive study. These compositions were chosen so as to represent five consecutive groups of quality which we were pleased to call "(1) Superior, (2) Good, (3) Average, (4) Fair, (5) Inferior." These five categories must be interpreted in terms of the available compositions, and must apply by no means to any fixed value which people may attach in general to these descriptive terms.

For the "Superior" group the ten compositions having the highest quality on the Hudelson English Scale were

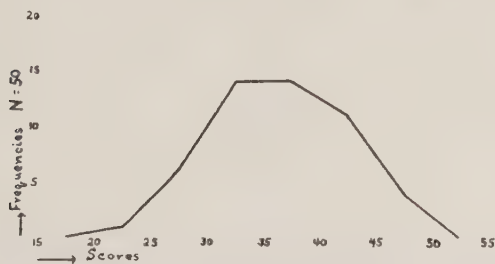


Figure 9 - Quality of 50 Specimens Chosen to Study Improvement Due to Training .

chosen. Similarly, the "Inferior" group consisted of the ten compositions having the lowest quality on this same scale. The range of quality between the limits of the two above groups was divided into three approximately equal parts, and ten compositions were chosen from each to represent qualities "Good," "Average," and "Fair," respectively, and to insure a random sample. The curve plotted from the qualities of these compositions, as given in Figure 9, is fairly normal. These represent also the five types of discourse in fairly equal proportions.

The average quality of each of these five groups on each of the seven composition scales after training is given in Table 8. The average improvement on each scale corresponding to each of these average qualities is also given in

the same table. This was calculated from a table giving the average improvement in the coefficient of variation of each composition as a result of training. The reader will find this table in the appendix. It will be noted that the assumption concerning the consistency of the quality of a composition irrespective of the scale upon which it is rated is justified by these data. The composite average improvement for each group on all scales is given in Table 8, and shows that the extremes of the curve improve most in objectivity as a result of training. The compositions of the "Average" and "Fair" qualities gain the least. This indicates that judges need more training to mark very good and very poor compositions objectively than those of average and mediocre quality. Again, it is the extremes that vary.

TABLE 8

IMPROVEMENT OF 50 SELECTED COMPOSITIONS REPRESENTING FIVE QUALITIES AND THEIR CORRESPONDING IMPROVEMENT IN TERMS OF COEFFICIENT OF VARIATION

" SUPERIOR " QUALITY								
	H. E. ³¹	H. T.	Leo.	Lewis	Nass.	V. W.	W.(S)	M.A.I.
No. of comps.	10	10	10	7	10	8	10	
Av. quality	4.48	4.02	5.06	44.26	5.07	44.11	50.24	
Av. improve.	6.96	8.23	4.30	4.99	3.22	6.20	6.09	5.71
" GOOD " QUALITY								
No. of comps.	10	10	10	7	10	8	10	
Av. quality	4.02	3.62	4.59	40.90	4.49	39.01	44.73	
Av. improve.	3.82	7.00	5.36	5.95	— .03	4.45	4.44	4.43
" AVERAGE " QUALITY								
No. of comps.	10	10	10	5	10	8	10	
Av. quality	3.62	3.24	4.20	37.56	4.07	35.26	40.88	
Av. improve.	2.43	7.97	4.86	5.46	1.44	3.71	— .14	3.68
" FAIR " QUALITY								
No. of comps.	10	10	10	5	10	9	10	
Av. quality	3.25	2.84	3.65	34.53	3.48	29.75	35.44	
Av. improve.	.60	12.23	6.58	2.04	.88	3.12	— 3.83	3.09
" INFERIOR " QUALITY								
No. of comps.	10	10	10	5	10	8	10	
Av. quality	2.85	2.49	3.18	27.49	3.02	25.16	31.95	
Av. improve.	.49	11.58	9.28	2.71	1.56	2.21	4.92	4.68

³¹ H. E.—Hudelson English. H. T.—Hudelson Typical. Leo.—Leonard. Nass.—Nassau. V. W.—Van Wagenen. W. (S)—Willing (Story). M. A. I.—Mean Average Improvement.

This same problem was studied from another angle. All data referring to this are given in Table 9. In this study the

gains in the coefficient of variation of the fifty selected specimens were ranked in size on each of the seven scales, and divided into five consecutive groups, as indicated in the table. The reader should understand that the same compositions do not compose each of the groups in each scale as do those in Table 8. The basis determining the group to which the compositions belong was the improvement in the coefficient of variation as a result of training for each scale separately. It will be noted that in practically every scale the compositions which make the most improvement have the highest quality; and that, in most cases, those compositions which make the least improvement (here actually lose in objectivity) represent those of the lowest quality. Two notable exceptions are the Hudelson Typical and the Leonard Scales. Here the compositions which gain the most in objectivity are those of the poorest quality, while in the Hudelson Typical the less gain there is the higher is the quality.

TABLE 9

IMPROVEMENT OF 50 SELECTED COMPOSITIONS REPRESENTING FIVE IMPROVEMENTS AND THEIR CORRESPONDING QUALITIES IN EACH OF SEVEN SCALES

GROUP I							
	H. E. ²²	H. T.	Leo.	Lewis	Nass.	V. W.	W. (S)
No. of comps.	10	10	10	6	10	8	10
Av. improve.	9.82	17.59	15.74	11.47	8.91	11.16	8.06
Av. quality	3.90	2.90	3.78	38.84	4.20	37.75	47.63
GROUP II							
No. of comps.	10	10	10	6	10	8	10
Av. improve.	5.44	11.53	8.84	7.66	3.09	6.17	4.51
Av. quality	3.70	3.07	4.34	35.24	3.73	31.95	44.98
GROUP III							
No. of comps.	10	10	10	6	10	8	10
Av. improve.	1.96	9.23	6.03	4.32	.91	3.50	—70
Av. quality	3.77	3.34	4.37	42.08	4.09	36.11	39.40
GROUP IV							
No. of comps.	10	10	10	6	10	9	10
Av. improve.	—03	6.53	2.60	.48	—1.31	1.66	—2.96
Av. quality	3.72	3.46	4.12	36.56	4.10	36.59	38.24
GROUP V							
No. of comps.	10	10	10	5	10	8	10
Av. improve.	—2.89	2.13	—2.82	—3.21	—4.52	—2.63	—8.67
Av. quality	3.11	3.43	4.07	35.54	4.00	30.03	33.01

²² For explanation of abbreviations see footnote number 31.

This seems to indicate that judges need training to mark compositions of inferior quality on the Hudelson Typical and the Leonard Scales, while the remaining scales demand more training for the objective use of the higher ends of the scales.

OBJECTIVITY OF SCALES AS DETERMINED BY CRITERION II

The derivation of Criterion II was explained in Chapter III. The forty compositions chosen represent an approximately normal distribution of quality as shown in Figure 10.

All that remains to be added at this point is a table of the ranks (according to quality) which these forty selected specimens take on the six scales (Lewis being omitted for reasons given in Chapter III) ; the criterion rank, and *rho* for each scale.

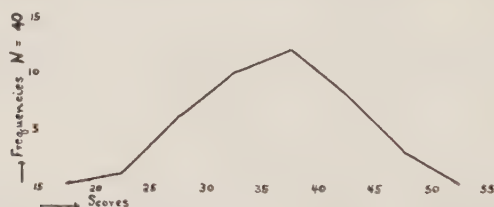


Figure 10 - Quality of 40 Specimens Chosen to Build up Criterion II.

Taking Criterion II as a measure of objectivity, the above evidence leads to the conclusion that all of the scales measure the same compositions very consistently, indeed. In other words, these six scales yield very objective results when rank of quality rather than actual quality is considered. The correlations are so high and so close that it was necessary to carry out *rho* to four places in order to effect a ranking of the scales. The corresponding Pearson coefficients of correlation are so high that the reliability of the results can scarcely be questioned.

If the scales are ranked in order of merit according to the size of *rho*, they would appear in the following order:

1. Van Wagenen, 2. Hudelson Typical, 3. Willing—Story Value, 4. Nassau, 5. Hudelson English, 6. Leonard.

The reader must be careful to avoid the conclusion that this order upsets that derived from application of Criterion I.

TABLE 10
BUILDING UP OF CRITERION II

Comp. ³³ Code	Rank on						Total Rank	C. II
	H. E.	H. T.	L.	N.	V. W.	W.		
21	1	1	1	1	1	1	6	1
59	2	2	2	2	3	2	13	2
98	3	4	15	6	4	11	43	5
60	4	6	4	4	7	4	29	4
90	5	8	14	8	10.5	15	60.5	10
2	6	4	3	5	2	3	23	3
94	7	4	10.5	12	5	7	45.5	6
3	8	11	8.5	3	8	14	52.5	9
99	9	9.5	16	10	9	9	62.5	11
58	10	7	6.5	7	10.5	6	47	7
71	11	12	21	14	16	20	94	17
37	12	14	19.5	9	14	21	89.5	16
13	13	9.5	5	13	6	5	51.5	8
67	14	16	10.5	17	15	16	88.5	15
19	15	15	8.5	11	12	8	69.5	12
62	16	20	12	23	19	12	102	18
36	17	13	13	16	13	10	82	13
64	18	17	6.5	15	17	13	86.5	14
55	19	19	17.5	19	20	18	112.5	19
52	20	22.5	19.5	22	23	19	126	21
11	21	24	26	26	24	24	145	25
31	22	27	27.5	18	22	26	142.5	24
22	23	21	22	24	21	23	134	22
1	24	22.5	17.5	20	18	17	119	20
66	25	18	24.5	21	27	25	140.5	23
28	26	29	36	29	33	34	187	31.5
61	27	25	24.5	25	26	27	154.5	26
35	28	29	32.5	30	30	30.5	180	29
10	29	31	34	32	29	30.5	185.5	30
8	30	26	29.5	27	25	28	165.5	28
49	31	32.5	27.5	31	36	29	187	31.5
54	32	32.5	32.5	33	34	33	197	33
20	33	35.5	29.5	35	35	32	200	34
6	34	29	23	28	28	22	164	27
57	35	35.5	37	37	38	38	220.5	37
72	36	37.5	31	36	32	35	207.5	35
45	37	37.5	40	38	39	39	230.5	39
18	38	34	35	34	31	36	208	36
51	39	39	38	39	37	37	229	38
12	40	40	39	40	40	40	239	40

Rho, or the rank coefficient of correlation between Criterion II and:

	<i>Rho</i>	Pearson <i>r</i>
Hudelson English	.9700	.97
Hudelson Typical	.9856	.99
Leonard	.9555	.96
Nassau	.9744	.97
Van Wagenen	.9857	.99
Willing	.9748	.98

Both criteria measure objectivity, it is true; but Criterion I measures absolute merit, while Criterion II measures relative merit. Furthermore, the rank coefficients of correlation are so close that the differences are, for practical purposes, negligible. In the light of these results it seems justifiable to conclude that, if order of merit is desired, it matters little which of the six scales is applied, provided the judge is trained in the use of such scales.

³³ The first vertical column of figures indicates the code numbers of the forty compositions selected from Table 5. H. E.—Hudelson English. H. T.—Hudelson Typical. L.—Leonard. N.—Nassau. V. W.—Van Wagenen. W.—Willing. C. II—Criterion II.

OBJECTIVITY OF SCALES AS DETERMINED BY CRITERION III

The nature and value of Criterion III for purposes of this study were discussed in Chapter III. Added to what has already been said there, the reader must bear in mind the fact that here scales are compared on the basis of their variability (in terms of standard deviation) only, and *not* on their variability in proportion to their central tendency, as in the case of Criterion I. This will account partly for apparent contradictions in rankings made on the basis of Criterion III. The two are *not* directly comparable, but throw light on the objectivity of the scales from different angles. Results are given in Table II.

TABLE 11 ³⁴ADVANTAGE IN OBJECTIVITY OF MARKS ASSIGNED BY USE OF
COMPOSITION SCALES OVER CHANCE

Scale	Before training				After training		
	Chance S. D.	Scale S. D.	Per cent	% bet. chance	Scale S. D.	Per cent	% bet. chance
Hud. Eng.	.97	.77	79.4	20.6	.76	78.4	21.6
Hud. Typ.	1.50	.98	65.3	34.7	.86	57.3	42.7
Leonard	1.25	.92	73.6	26.4	.92	73.6	26.4
Lewis	12.78	10.01	78.3	21.7	9.68	75.7	24.3
Nassau	1.58	.88	55.7	44.3	.97	61.4	38.6
Van Wag.	11.28	8.92	79.1	20.9	8.76	77.7	22.3
Willing	13.23	10.55	79.7	20.3	11.07	83.7	16.3

Each of the standard deviations is in terms of the particular scale units of the respective scales. The superiority of one scale over another in objectivity on this basis depends on the per cent which it is better than chance, the one having the highest percentage being the best.

If the scales are thus ranked, the following decreasing order of merit will obtain before training:

1. Nassau, 2. Hudelson Typical, 3. Leonard, 4. Lewis, 5. Van Wagenen, 6. Hudelson English, 7. Willing.

³⁴ This table should be read: On the Hudelson English Scale before training the standard deviation of marks assigned by chance is .97. The average standard deviation of the hundred compositions marked on this scale is .77. This is 79.4 per cent of the chance standard deviation, and it is, therefore, 20.6 per cent better (smaller) than that of chance ratings; etc.

The last three are very closely equal in their objectivity as measured by this criterion.

After training only two scales change position; namely, Hudelson Typical, which takes first place, and Nassau, which assumes second place. This again emphasizes the merit of the Hudelson Typical Scale in the hands of trained judges. The order is:

1. Hudelson Typical, 2. Nassau, 3. Leonard, 4. Lewis, 5. Van Wagenen, 6. Hudelson English, 7. Willing.

Relatively, the variability of the scales remains practically unchanged after training. Absolutely, however, five of the scales improve in objectivity (decrease in variability) as a result of training. (See Table 11.) As shown all through the study, training reduces variability in almost every case, and, hence, is a necessary preparation for the competent use of composition scales.

OBJECTIVITY OF DISCOURSE DISTINCTION SCALES VS. NON-DISCOURSE DISTINCTION SCALES

The non-discourse distinction scales are the Hudelson English, Hudelson Typical, Leonard, Nassau, and the Willing. The Lewis and the Van Wagenen make discourse distinctions. In Table 5 the arithmetic average of the means, standard deviations, and coefficients of variation were given for each of the five groups of compositions which exemplify four types of discourse and two types of writing as explained in Chapter II. These results, arranged in a form better suited to the present purpose, are given in Table 12.

Inspection of the coefficients of variation for each of these groups will show that the non-discourse distinction scales, almost without exception, measure the various types of discourse more objectively than the discourse distinction scales. The one consistent exception to this rule is the Hudelson Typical Scale, which has a great variability before training as previously brought out.

TABLE 12

ARITHMETIC AVERAGE OF THE MEANS, STANDARD DEVIATIONS, AND
COEFFICIENTS OF VARIATION FOR FIVE GROUPS OF COMPOSITIONS
EXEMPLIFYING DIFFERENT TYPES OF DISCOURSE

BEFORE TRAINING ³⁵					
	Des.	Exp.	Nar.	Arg.	O. L.
<i>Hudelson English</i>					
M.	3.14	2.96	3.30	3.41	3.56
S. D.	.72	.70	.75	.80	1.01
C. V.	22.98	23.80	22.94	23.59	28.13
<i>Hudelson Typical</i>					
M.	2.69	2.48	2.79	2.97	3.26
S. D.	.99	.91	.94	1.01	1.13
C. V.	38.02	37.27	34.58	34.77	35.04
<i>Leonard</i>					
M.	3.63	3.10	3.56	3.61	3.30
S. D.	.96	.86	.93	.91	.93
C. V.	27.21	28.61	27.51	25.86	28.61
<i>Lewis</i>					
M.			33.05	33.14	35.23
S. D.			9.02	10.14	11.68
C. V.			28.40	31.07	33.36
<i>Nassau</i>					
M.	3.35	3.02	3.39	3.63	3.73
S. D.	.86	.83	.84	.91	1.03
C. V.	26.06	27.58	25.55	25.15	27.71
<i>Van Wagenen</i>					
M.	20.64	25.53	20.55		33.91
S. D.	8.44	7.59	8.48		13.09
C. V.	29.30	30.20	28.93		38.68
<i>Willing</i>					
M.	36.06	33.90	36.78	36.66	35.75
S. D.	9.36	10.03	9.60	11.06	14.91
C. V.	26.11	30.85	26.30	30.39	41.59
AFTER TRAINING					
<i>Hudelson English</i>					
M.	3.56	3.36	3.68	3.82	4.02
S. D.	.70	.75	.73	.82	.87
C. V.	19.97	22.30	19.97	22.03	21.78
<i>Hudelson Typical</i>					
M.	3.20	2.96	3.25	3.39	3.61
S. D.	.83	.85	.84	.87	.99
C. V.	26.44	28.80	26.24	26.22	27.42
<i>Leonard</i>					
M.	4.26	3.75	4.24	4.25	4.09
S. D.	.87	.87	.86	.92	1.26
C. V.	20.96	23.55	20.80	22.38	31.00
<i>Lewis</i>					
M.			39.10	36.41	39.61
S. D.			9.39	7.78	12.42
C. V.			25.13	22.07	31.72
<i>Nassau</i>					
M.	4.03	3.74	4.05	4.21	4.33
S. D.	.95	.95	.95	.98	1.09
C. V.	24.10	25.74	23.71	23.46	25.40
<i>Van Wagenen</i>					
M.	36.06	30.48	34.25		37.98
S. D.	8.06	8.21	7.96		12.57
C. V.	23.07	27.34	23.84		33.21
<i>Willing</i>					
M.	41.38	36.63	41.83	41.30	41.38
S. D.	10.56	10.64	10.32	11.23	14.08
C. V.	26.29	29.23	25.38	27.99	34.02

³⁵ For explanation of all abbreviations see footnote number 25. M = mean.

Conclusions cannot be drawn so readily from the data after training as from those before training. For this reason the arithmetic average of the coefficients of variation on the discourse distinction scales was calculated and compared with the non-discourse distinction scales for each of the five types of discourse. For the sake of completeness, the same was done before training, and results are given in Table 13.

TABLE 13

AVERAGE COEFFICIENTS OF VARIATION OF DISCOURSE AND NON-DISCOURSE DISTINCTION SCALES

BEFORE TRAINING					
	Des.	Exp.	Nar.	Arg.	O. L.
Non-discourse Distinction scales	28.08	29.62	27.38	22.92	32.22
Discourse Distinction scales	29.30	30.20	28.67	31.07	36.02
AFTER TRAINING					
Non-discourse Distinction scales	23.55	25.92	23.22	24.42	27.92
Discourse Distinction scales	23.07	27.34	24.49	22.07	32.47
Average of non-discourse Distinction scales is		25.01			
Average of discourse Distinction scales is		25.89			

Only by taking the composite average of each of the two types of scales on all discourses as a whole is it possible to show that there is a very slight difference in favor of the non-discourse distinction scales. This is practically negligible. This seems to indicate that judges must be trained in order to mark objectively on discourse distinction scales. From the standpoint of objectivity alone, both types are probably equally objective in the hands of trained judges. Other arguments in favor of a discourse distinction scale might be raised (as will be shown later in the opinion of the judges in this experiment) which make that type of scale valuable; but if viewed from the standpoint of objectivity and the time required to learn and to use such a scale, the non-discourse distinction scales will probably prove to be the more satisfactory.

GENERAL MERIT VS. DIAGNOSTIC SCALES

The two diagnostic scales which figure in this experiment are the Van Wagenen and the Willing. By this type of scale we mean one which assigns separate values to such items as thought, structure, and mechanics. The Van Wagenen gives a separate score for each of these three aspects and combines them into a general merit rating in which thought is weighted four, structure two, and mechanics one. Only the general merit rating was considered in all of the mathematical computations of this study. Separate scores for the three elements are in the writer's possession, but have not yet been considered separately. She feels, however, that the diagnostic values are not lost in the composite score because of the weighting system which enters into its composition. Some time in the near future she expects to consider these elements separately, and, hence, the conclusions herein drawn must be taken as tentative.

The Willing Scale divides a composition into two phases, and assigns one value for story merit and a totally different one for form. The two are not capable of direct combination³⁸ into one composite score, because the form value is in terms of the number of grammatical and other form errors per hundred words.

The Leonard Scale approaches somewhat the diagnostic scales in character in that it disregards completely matters of form, for which it provides no system of scoring whatsoever, and considers composition quality or thought only. For purposes of this comparison, however, we will designate the Van Wagenen and the Willing as the two diagnostic scales (since they conform to the orthodox definition of such a scale), and the rest as general merit scales.

³⁸ See Willing, M. H. "Willing's Scale for Measuring Written Composition." Public School Publishing Co., Bloomington, Ill. Also, School Survey of Denver, Part II, "The Work of the Schools"—The School Survey Committee, Denver, Col., 1916, p. 61.

TABLE 14
AVERAGE COEFFICIENTS OF VARIATION OF DIAGNOSTIC AND GENERAL
MERIT SCALES

	BEFORE TRAINING					
	Des.	Exp.	Nar.	Arg.	O. L.	Aver.
Diagnostic	27.71	30.53	27.62	30.39	40.14	31.28
General merit	28.57	29.32	27.80	28.09	30.57	28.87

AFTER TRAINING						
Diagnostic	24.68	28.29	24.61	27.99	33.62	27.84
General merit	22.87	25.10	23.17	23.24	27.46	24.37

These averages indicate that general merit scales measure given compositions more objectively than the diagnostic scales we now have, both in the hands of trained and untrained judges. A further disadvantage of a diagnostic scale is the amount of time consumed in employing it as compared with that on the general merit scale for the same compositions. Evidence for this statement will be given under the heading of "Administrative Feasibility of the Scales" later in this chapter. The use of diagnostic scales must be justified on grounds other than objectivity and practicability. The conclusion of Briggs,³⁷ ". . . the critics of general merit scales have not, at least in published reports, accepted these substitutes (diagnostic and discourse distinction scales). They are useful in their way, but they cannot replace the general merit type for measuring the result of marking," seems justifiable.

EFFECT OF TRAINING ON HILLEGAS AND NON-HILLEGAS DERIVATIVES

The reader will recall that in Chapter I a Hillegas derivative was defined as a scale which is either a modification of the original Hillegas Scale or one which is evaluated in Hillegas units. Four of the seven scales are of the former type. These are the Nassau Supplement to the Hillegas Scale, Hudelson's Virginia Supplement to the Hillegas Scale (Hudelson English), Hudelson's Typical Composition Abil-

³⁷ Briggs, Thomas H. "English Composition Scales in Use." *Teachers College Record*, Vol. 23, November, 1922, p. 428.

ity Scale, and the Lewis English Scales for Letters. The three which exemplify the other variety are the Van Wagenen English Composition Scale, the Willing English Composition Scale, and the Leonard Scale for the Measurement of Composition Quality Only.

The average of the coefficients of variation of the two groups, before and after training, are given in Table 15.

TABLE 15
AVERAGE COEFFICIENT OF VARIATION OF HILLEGAS AND NON-HILLEGAS
DERIVATIVES

	Before training	After training	Gain
Hillegas derivatives	29.22	24.62	4.60
Non-Hillegas derivatives	29.48	25.71	3.77

As was expected, the Hillegas derivatives improve most in objectivity as a result of training on Hillegas practice material, such as was described in Chapter I. It should be noted, however, that the non-Hillegas derivatives improve considerably from an application of the same material. The difference in the amount of the improvement of the two types of scales is sufficient, it seems, to indicate the desirability of material specially adapted to the type of scale upon which a judge may want to train himself. In the absence of this, the scorer can effect appreciable improvement on the material now available. This statement might prove to be still more significant in view of the fact that the Leonard coefficient of variation (Table 6) before training may be greater than it would ordinarily be due to the fact that it is based on nine judgments only. (That after training is based on twenty-five judgments, just as all the other scales.)

EFFECT OF SPECIAL FACTORS OF SCALES ON THEIR OBJECTIVITY

Number of scale-steps. The scales vary in the number of steps into which they are divided from 8 on the Willing to 16 on the Hudelson English, the most common number being 10 (Hudelson Typical, Leonard, and Nassau). This feature of each scale has been discussed under "Scales Used," in

Chapter I. The reader will recall that the Van Wagenen has fourteen, and the Lewis nine, such steps, on the average, for each of its component scales. Those scales which have less than ten steps will, for purposes of this comparison, be grouped together and designated as "short" scales, with Willing and Lewis composing this class. Those having ten will constitute the second, or "average," group, and will consist of the Hudelson Typical, Leonard, and the Nassau. Those having more than ten steps fall in the last group, designated as "long," and comprise the Hudelson English and the Van Wagenen. For results see Table 16.

TABLE 16
AVERAGE COEFFICIENTS OF VARIATION OF "SHORT,"
"AVERAGE," AND "LONG" SCALES

	Before training	After training
Short	30.18	26.97
Average	30.04	24.86
Long	27.41	23.56

From these results it seems reasonable to conclude that, irrespective of training, the longer the scale, the more objective the scores. This must not be construed to mean that length in and of itself determines the objectivity of a scale, but of the three types represented in this study data do seem to point to the conclusion that a longer scale gives the rater a larger number of samples and a finer gradation of specimens against which to match his particular compositions, and thus to determine more accurately their real merit.

Training reduces the coefficient of variation for each of the three groups consistently, but does not seem to affect the relative objectivity of the three groups. Hence, both the trained and the untrained judge will do better to choose a scale which offers him a larger variety of specimens for comparison.

Statement of merits and defects. The judges participating in this experiment were loud in their praises of those scales in which each sample is followed by a statement of its merits and defects, which constitutes a sort of justification for its

position in the scale. Dolch³⁸ holds that such a statement would help to obviate three very distinct sources of inaccuracy in the use of English composition scales. The only scales which have such comments are the Leonard and the Harvard-Newton. The latter cannot be considered for reasons explained in Chapter I; hence, the only data available for making a comparison between scales with and without this feature are very scant. Further, the small number of judgments upon which the Leonard coefficients of variation *before* training are based make such comparisons still more unreliable. Hence, the figures given in Table 17 must be considered very tentatively.

TABLE 17

AVERAGE COEFFICIENTS OF VARIATION OF SCALES WITH AND WITHOUT
A STATEMENT OF MERITS AND DEFECTS FOLLOWING
EACH SCALE SAMPLE

	Before training	After training
Leonard	27.60	23.06
Other scales	29.62	25.43

Even with all of the limitations described above, Table 17 shows coefficients of variation which are smaller on that scale in which each sample is followed by a statement of its merits and defects than all the other scales lumped together. It might not be unreasonable to conclude that our judges were right in their opinion of the helpfulness of such statements in yielding standards by which to judge a composition.

Number of samples per scale-step. Some constructors of scales seem to feel that a judge can make a more efficient judgment of the real merit of a composition if he is given several samples of approximately equal value on each scale-step. Proceeding on this assumption Thorndike, Leonard, and Hudelson, in his Typical Composition Ability Scale, have provided for this. The samples become more numerous on the scale-steps in the middle of the scale and decrease as they approach the extremes. The question arises: "Does

³⁸ Dolch, E. W. "More Accurate Use of Composition Scales"—English Journal, Vol. 11, November, 1922, pp. 536-544.

this feature make for more uniformity of judgment than the provision of only one sample per scale-step?" Let us inspect the average coefficients of variation of the scales of these two types.

TABLE 18

AVERAGE COEFFICIENTS OF VARIATION FOR SCALES CONTAINING ONE SAMPLE PER SCALE-STEP (GROUP I) AND SCALES CONTAINING MORE THAN ONE SAMPLE PER SCALE-STEP (GROUP II)

	Before training	After training
Group I	28.31	25.10
Group II	31.87	25.06

In view of what has been said repeatedly concerning the Leonard coefficient of variation before training, and of the necessity for training for a satisfactory use of the Hudelson Typical Scale, results in Table 18 might not be misinterpreted if one concludes that several samples per scale-step are an aid to the trained scorer. Considered from the standpoint of training alone, both types of scales can be used with practically equal promise of success in uniform marking. The judges participating in this experiment, however, seemed to think that a variety is quite an aid.

SPECIAL ADAPTABILITY OF THE LEWIS SCALE FOR LETTERS

The Lewis Scales are designed especially for letter writing, the most common and universal type of writing which adults must perform. It is important that a scale be available to rate this kind of writing. The question arises: "Does the Lewis do this better than the other scales, so far as objectivity is concerned?" To answer this consult Table 19.

TABLE 19

AVERAGE COEFFICIENT OF VARIATION OF LETTER COMPOSITIONS ON LEWIS VS. OTHER SCALES

	Before training			After training		
	Nar. So. ³⁰	Exp. So.	Order L.	Nar. So.	Exp. So.	Order L.
Lewis	28.40	31.07	33.36	25.13	22.07	31.72
Other scales	27.64	27.95	33.29	23.32	24.42	28.81

³⁰ Nar. So.—Narrative Social Letter, otherwise designated as Narration. Exp. So.—Expository Social, otherwise designated as Argumentation. Order L.—Order Letter.

The differences in variability are not very great, but they lead to the tentative conclusion that the Lewis Scales do not evaluate letters as objectively as the other scales do. This is true both before and after training, except in the case of the expository social letters. Training does reduce the variability of the Lewis Scale considerably, but even trained scorers cannot expect to mark letters more uniformly by employing this scale than they could by employing any other.

SUBJECTIVE RANKING OF SCALES BY JUDGES IN ORDER OF PREFERENCE

For an explanation of this procedure see Chapter II. A table presenting results of the judges' preferences follows:

TABLE 20
RANKS OF SCALES IN ORDER OF JUDGES' PREFERENCE

Scale	Median rank
Van Wagenen	1.86
Leonard	3.3
Hudelson English	3.5
Nassau	5
Hudelson Typical	5
Lewis	5.3
Willing	7.2

The most decisive results were obtained in the case of the Van Wagenen and the Willing Scales. Fourteen judges ranked the former first, and fourteen ranked the latter last. On no other scale was there such uniformity of agreement. The reasons given by the judges for their overwhelming preference for the Van Wagenen Scale were:

1. "It gives the proper proportionate values to all the elements which make up a composition."
2. "It helps a teacher most in discovering the needs of her pupils."
3. "The gradation in scale-steps is fine enough to accommodate all types of compositions."
4. "It gives more value to *thought*, which I consider most important."
5. "It helps the marker most in determining the exact place in the scale where a given composition belongs."
6. "It provides a scale for each type of writing (*e. g.*, narration, exposition, description)."
7. "It suits these sixth grade compositions best."
8. "It gives one general mark for the composition as a whole."
9. "It is fairest to the child."

These reasons are stated in the order of their frequency, number "1" having been mentioned most often as a justification for the judges' preference. Evidently the diagnostic value of the scale appeals most to these judges.

The reasons given by the judges for their decided rejection of the Willing Scale were:

1. "It takes too much time to mark."
2. "It is the most complicated scale to use."
3. "It does not give one general mark for the composition as a whole."
4. "The gradation in scale steps is not fine enough to accommodate all types of compositions."
5. "It does not give the proper proportionate weight to 'Form' and 'Story' values."
6. "It does not provide a separate scale for each type of discourse."
7. "It is difficult for the marker to determine the exact place in the scale where a given composition belongs."
8. "It does not suit these sixth grade compositions."

The above reasons are stated in the order of frequency with which they were mentioned by the judges, number "1" having been mentioned most often. To present the chief objection succinctly which judges stated vigorously both on their score sheets and repeatedly to the writer in person, the latter will cite the following quotation: "The 'Story Value' part of the scale is entirely satisfactory, but the counting of words is too impracticable. Today we are concerned with speed and accuracy. We are to lighten the burdens of teachers, but this scale is an added burden." A good deal of irritation was also felt because there is no composite score for a composition.

If the ranks of the scales in Table 20 are compared with those resulting from an application of Criterion I after training, it is apparent that there is little agreement. The only case in which there is a one to one correspondence is the Willing Scale, which ranks last according to Criteria I, III, and the preferences of the judges. This comparison proves that opinion, even of trained raters, cannot be relied upon to determine the absolute merit of composition scales. At least, opinion is no measure of the objectivity of a scale. Neither

opinion nor experience can act as substitutes for experimentation.

ADMINISTRATIVE FEASIBILITY

Time. A consideration which no educational test, method, or device can escape is its practicability. The first question which the administrator and over-worked teacher asks is: "How much time will it take, and what will I get in return?" For this reason the writer is attempting to ascertain which scale gives the most objective results in the least time without training, and which after training.

TABLE 21
AVERAGE TIME PER JUDGE PER COMPOSITION

	Before training min.	After training min.
Hudelson English	.83	.82
Hudelson Typical	.89	.75
Leonard	.84	.77
Lewis	1.15	1.11
Nassau	.78	.67
Van Wagenen	2.44	1.95
Willing	2.57	1.82

The figures in Table 21 were calculated by taking the average time which it took a judge to mark the whole set of compositions and then dividing by the number of compositions in the set. In the case of all scales, except the Lewis and the Van Wagenen, this number was one hundred. On the latter scales the number was fifty-two and eighty-five, respectively, as previously explained. The time it took a judge to become familiar with the scale is not included in these results. Table 21 considers only the actual average time consumed to mark one composition (each of which was only one paragraph in length) on each of the respective scales. It is clear from these data that the scale which yields the most objective results in the shortest time, both before and after training, is the Nassau Supplement to the Hillegas Scale.

If the scales are ranked in the order of the average time necessary to mark a composition, they would stand in the following order :

Scale	Rank before training	Rank after training
Nassau	1	1
Hudelson English	2	4
Leonard	3	3
Hudelson Typical	4	2
Lewis	5	5
Van Wagenen	6	7
Willing	7	6

The only appreciable shift due to training is that of the Hudelson English Scale. It will be noticed that training makes little difference in the time taken to mark compositions on this scale. In view of the fact that this is a very objective scale in the hands of untrained markers, in addition to being one which takes a relatively short time to use, it is to be recommended especially to people untrained in the use of composition scales. The best scale, however, for both trained and untrained scorers is the Nassau Supplement.

Trained judges can rate more rapidly on every scale than untrained judges. This is rather to be expected.

Both of the diagnostic scales take the longest to use. Administrative feasibility cannot be urged for these, but the values of such a scale as the Van Wagenen, for instance, as stated by the judges, adds considerable significance to its value.

Grade suitability. Administrators everywhere must answer the question: "What suits a given grade the best?" For the alleged grade suitability of each of the scales the reader is referred to the bibliography. In order to determine which of these scales seems to fit these sixth grade compositions best, a study was made of the number of undistributed scores in each scale after training. By this is meant the number of scores which fall on the lowest and on the highest steps of the respective scales. The theory underlying this is the contention that if a composition falls on the lowest step, it is impossible to measure how much lower it would score if

lower steps were provided, likewise with the high steps. None of the marks on our compositions reached the highest limit of any of the scales; hence, all undistributed scores will refer to the number which fell on the lowest step. The number of these (after training only were considered) and their corresponding per cents are given in Table 22.

TABLE 22
UNDISTRIBUTED SCORES AFTER TRAINING

	Total	Scores	Per cent
Hudelson English	2500	129	5.1
Hudelson Typical	2500	0	0
Leonard	2500	40	1.6
Lewis	1300	48	3.7
Nassau	2500	0	0
Van Wagenen	2125	2	.1
Willing	2500	202	8.1

In the light of these data the Nassau and Hudelson Typical Scales suit the sixth grade compositions best. The Van Wagenen also is fitted for this grade. Leonard can be used practically without error, and the Willing and Hudelson English can be applied without very serious error, but are better suited to grades above the sixth. Considering Criteria I and III, the administrative feasibility, and the grade suitability, the Nassau Scale is the best.

USE OF SCALES IN SCHOOL SURVEYS

The writer examined seventy-eight surveys covering a period of seventeen years, 1912 to 1928, inclusive. This search has revealed the fact that very few surveys have employed composition scales in the past, but that they are increasingly used as the present time is approached. Of all the scales available five were used in these seventy-eight surveys. The most often used is the Nassau County Supplement to the Hillegas Scale. The next most frequently used is the Hillegas Scale for the Measurement of English Composition by Young People. The composition scale used, the

place, the grades in which such was used, and the date of the surveys, are as follows:

Harvard-Newton English Composition Scale		Date
Brookline	Grades 6- 8	1917
Missouri	" 7- 8	1922
Port Townsend (Wash.).....	" 4-12	1915
Hillegas Scale for the Measurement of English Composition by Young People		
Salt Lake City.....	Grades 4- 8	1915
Butte	" 4- 8	1914
Janesville	" 3-10	1918
Nassau County	" 4-12	1917
Gary	" 4-12	1918
Ohio	" 8	1914
Hudelson English Composition Scale		
Philadelphia	Grades 8-12	1922
Nassau County Supplement to the Hillegas Scale		
St. Paul	Grades 4-12	1917
Binghamton	" 4- 8	1919
Utica	" 4- 8	1920
Baltimore	" 4-12	1920
Virginia	" 11	1921
Niagara Falls	" 4- 9	1921
Lawrence Township	" 4- 8	1922
Stamford	" 4- 8	1923
Tampa	" 4-7-10-12	1926
Willing English Composition Scale		
Denver	Grades 4- 8	1916
Grand Rapids	" 4- 8	1916

The educational experts who conducted school surveys have used, for the most part, a very reliable scale; namely, the Nassau, and they have used it in grades to which it was adapted. It must be remembered, however, that the writer has evidence on the suitability of the scales for the sixth grade only.

The Harvard-Newton Scale has been used in three surveys, as indicated. It is most surprising to find that anyone conducting a survey should have chosen this scale for use in any grade lower than the eighth. How anyone could choose it for a grade as low as the fourth is inconceivable. The writer of the present study found that in the sixth grade this scale is too high for 61.6 per cent of the compositions marked. This means that only 38.4 per cent of the compositions written by normal sixth grade pupils are of sufficiently

high quality to fall within the Harvard-Newton range of values. Further evidence is presented in Appendix I. Those conducting surveys in Brookline and Missouri made neither criticisms nor recommendations on the basis of their findings, but those in Port Townsend did. They found that fourth grade compositions were, on the whole, poorer in quality than the lowest Harvard-Newton specimens, while medians for grades V, VI, and VII are equal to the lowest of the scale samples. The judges concluded:⁴⁰

1. "In mechanics—poor spelling, punctuation, sentence structure, and paragraphing."

2. "In thought—an almost total lack of any imaginative quality; a mathematical, mechanical treatment of the subject."

Such a conclusion is very unfair and depressing in its effect upon teachers, and is not warranted in the grades below VIII.

The Willing Scale which was used in the fourth, fifth, and sixth grades, as well as some others, is not a very objective scale, nor is it adapted to use in these grades. (See Table 22.) In view of this, Willing's statements⁴¹ that "classes in some schools write with far greater ease than the classes of other schools"; that in certain schools "greater attention needs to be given to written expression"; that these schools may well learn of those which do better; and that "supervisory assistance is necessary in a number of schools," lose their significance.

The use of the Hudelson English Composition Scale in grades 8-12 in Philadelphia is probably quite justified on the bases of objectivity and grade suitability.

⁴⁰ A Survey of the Port Townsend Public Schools. 1915 Bulletin of the University of Washington, University Extension Series No. 17, General Series No. 95, p. 101.

⁴¹ Denver Survey, Part II, 1916, pp. 63-64.

CHAPTER V

SUMMARY OF CONCLUSIONS

The writer will now answer the questions raised in Chapter I. These conclusions may be drawn subject to the conditions of this investigation and to the limitations of the statistical techniques employed, as fully set forth in connection with each individual problem in Chapters III and IV.

SPECIFIC CONCLUSIONS

1. When the scales are ranked in objectivity according to the three criteria, they take the following positions :

TABLE 23
SCALES RANKED ACCORDING TO CRITERIA I, II, III

Scale	Before training		After training		Crit. II ⁴²
	Crit. I	Crit. III	Crit. I	Crit. III	
Hudelson English	1	6	1	6	5
Hudelson Typical	7	2	6	1	2
Leonard	3	3	2	3	6
Lewis	5	4	4	4	..
Nassau	2	1	3	2	4
Van Wagenen	6	5	5	5	1
Willing	4	7	7	7	3

This shows, with only a single exception, a very close agreement in objectivity as measured by Criteria I and III, especially after training. The reader will recall that Criterion II measures objectivity in terms of relative rank only, and that the correlations are so close, that the scales are almost equally objective if relative merit of compositions alone is sought. The most objective scale for both trained and untrained judges is the Nassau Supplement to the Hillegas Scale. A close second is the Leonard for the Judgment of

⁴² Criterion II considers the ranks of scales only after training. See Chapter III.

Composition Quality Only. The least objective scale is the Willing Scale for Measuring Written Composition. This is especially true after training.

2. The compositions of "Superior" and those of "Inferior" quality (chosen on a five-fold basis of quality) are those which improve most in the uniformity of marks assigned to them as a result of training the judges. The three middle groups of intermediate quality gain less, and the nearer one approaches the "Average" group, the smaller is the amount of improvement in objectivity. Therefore, judges of composition need more training to score very good and very poor compositions objectively than those of average quality. Training is particularly necessary to grade themes of very poor quality by means of the Hudelson Typical Composition Ability and the Leonard Scales.

3. Persons untrained in the use of composition scales will find themselves varying much more in their judgment of the quality of themes by comparison with discourse distinction scales, such as the Van Wagenen and the Lewis, than by comparison with scales making no such divisions of writing into the various discourses. Scales of the latter type are the two Hudelsons, the Leonard, the Nassau and the Willing. Competent judges, however, will most probably render as objective judgments with one as with the other type of scales. Thus, discourse distinction scales will render objective results only in the hands of trained scorers.

4. Judges rating letters on the Lewis Scale, which is supposedly specially adapted to letters, do not, in general, render more objective judgments for such a specialized type of writing than they do on the other composition scales. This holds good for both trained and untrained judges.

5. Judges trained by the use of Hillegas practice material improve more in the uniformity of the judgments which they render on scales which are modifications of the original Hillegas Scale, or which are evaluated in its units (so-called "Hillegas derivatives"), than they do on "non-Hillegas derivatives." For this reason specially adapted practice series consisting of compositions of known value should be provided

for the Van Wagenen, the Willing, and the Leonard Scales. The first of these has such a series, but it is not sufficiently extensive to supply enough material for training to the point of competency (.5 points average deviation from expert judgment). Persons may train for the other scales on the Thorndike material used in the present study. It should be noted, however, that considerable improvement in uniformity of rating can be effected on any scale by practicing on the Thorndike series.

6. Both trained and untrained scorers rate more objectively on general merit scales (such as the Nassau, the two Hudelsons, etc.) than on diagnostic scales. The two diagnostic scales studied in this investigation were the Van Wagenen and the Willing. Of these the Van Wagenen is undoubtedly the more objective, statistically sound, and satisfactory (as judged by fulfillment of the purpose for which it was constructed) of the two. The most objective general merit scale is the Nassau Supplement. Because of the number of factors involved, and the complexity of the judgments to be made in diagnosing the various constituents of a composition, a great deal more careful training is necessary to achieve an equal degree of objectivity of scoring by diagnostic scales than by general merit scales.

7. Willing requires scorers of composition using his scale to assign a separate value to "Form" by counting the number of words in the composition, the number of errors in spelling, punctuation, and grammar, and calculating the number of "form errors" per hundred words. Besides being a very time consuming procedure, it is a highly subjective one, as shown by the results of this study given in the footnote number 29. This proves the absolute necessity of an external standard for obtaining objective results, even in a matter so generally accepted as being non-subjective as an obvious violation of a common grammar rule.

8. Results seem to indicate that, irrespective of training, a longer scale yields more objective results than short scales, or scales of average length. This does not mean that length of the scale is the sole determinant of its objectivity, but the

evidence gathered does seem to indicate that a longer scale gives the rater a larger number of samples and a finer gradation of specimens against which to match his particular compositions, and thus to determine more accurately their real merit.

9. A comparison of the ranks of the scales according to merit in the opinion of the judges, with the corresponding ranks according to Criteria I, II, and III shows little agreement. This shows that neither opinion nor experience in the use of composition scales can act as substitutes for experimentation in determining the most objective scale.

10. A statement of the analysis of the merits and defects following each sample on a composition scale, and acting as a sort of justification for the position of the sample in the scale, does seem to help the judge to identify the composition under consideration with the scale sample which it matches most nearly in quality, and, thus, causes more uniformity in judgment among the several judges rating a given set of compositions. In other words, a statement of the relative merits and defects of each sample in a composition scale makes for objectivity.

11. Administrators and busy English teachers must know which instrument requires the shortest time, and yields the best results. If the scales are ranked according to the average time consumed per composition, they stand in the following order:

	Before training	After training
Nassau	1	1
Hudelson English	2	4
Leonard	3	3
Hudelson Typical	4	2
Lewis	5	5
Van Wagenen	6	7
Willing	7	6

It is clear from this that the scale which gives the most objective results in the least time both before and after training is the Nassau Supplement to the Hillegas Scale. The Leonard Scale ranks second on these two bases. In the hands of trained scorers the Hudelson Typical Scale will yield

rather objective results in a relatively short time. The diagnostic scales consume the longest time, and their use must be justified on grounds other than administrative feasibility. As is to be expected, the length of time consumed in marking a set of compositions on any scale decreases with training and experience in their use.

12. A study of the per cent of undistributed scores in each scale after training leads to the conclusion that the scales which are best suited to these sixth grade compositions are the Nassau, the Hudelson Typical, the Van Wagenen, and the Leonard. The Lewis, the Hudelson English, and the Willing can be applied without very serious error, but are better suited to grades above the sixth. The Harvard-Newton, as shown in Appendix I, is not at all adapted for use in this grade.

13. By actual count, the majority of educational experts who have used composition scales in conducting surveys, have employed the most objective scale (the Nassau County Supplement to the Hillegas Scale), and have, in almost every case, applied scales which have a high degree of objectivity and which are adapted to the grades in which they were used. Three very serious exceptions, however, deserve special attention. These are the use of the Willing Scale for the fourth grade in Denver and Grand Rapids Surveys, and the employment of the Harvard-Newton Scales in the sixth grade in Brookline, Massachusetts, and, above all, in the fourth grade in Port Townsend, Washington, surveys. On the ground of the evidence presented in Chapter IV, this is unjustifiable, especially where recommendations followed a measurement of the results.

SCALE SUGGESTIONS TO ENGLISH TEACHERS AND SUPERVISORS

A busy English teacher of the sixth grade who is unfamiliar with composition scales, but who wishes to get a quick, objective measure of the general merit of the composition work of the pupils of her class, will get the best results from a use of the Nassau Supplement to the Hillegas Scale.

If such a teacher wishes merely to get the relative standing of her pupils with respect to their "all-round" achievement in composition work, perhaps for purposes of classification, it makes little difference which scale she uses so far as objectivity is concerned; but, in so far as time counts in relation to training, she will do well to choose the Nassau or the Hudelson English Scale.

The teacher who wishes to assess the efficacy of her endeavors to develop creative talent, and who, therefore, desires to measure the ideas only, finds a simple, practical, and reliable, well adapted instrument for the purpose in the Leonard Scale for the Judgment of Composition Quality Only.

If the teacher or supervisor is confronted with specially difficult cases of pupils who seem to remain the same, or even grow worse in their composition achievement in spite of all her teaching, she may be able to unravel the mystery and place her finger upon the source of the difficulty by applying the Van Wagenen English Composition Scales, provided she is willing to spend time to familiarize herself thoroughly with the scale and to train herself to a reasonable degree of accuracy. This is easily possible with the practice material which the author of these scales has provided in the scale manual itself.

The supervisor who wishes to induce her more or less reluctant teachers into using a composition scale will probably find that they have no objections at all to learning so simple, short, and usable an instrument as the Nassau County Supplement. Besides this, the evidence gathered in this experiment gives every reason to expect the most objective, quick, ready (without training), and easily interpreted results for general merit.

If the teacher feels the need of a special scale for different sorts of letters, she may turn to the Lewis Scales; but the advantage resulting from the type of writing which the samples herein represent is probably offset by the time consumed in marking and the relative objectivity of results in comparison with the general merit scales.

The teacher or supervisor who is anxious to measure the result of her instruction and supervision in terms of the amount of improvement between successive applications of the scale, should be careful not to be too hasty. She should follow the advice of Briggs (see footnote number 7), and apply the scale only once, or, at the most, twice a year. This caution should be observed because improvement in so complex a skill as composition writing is very slow, and too frequent measurement of results would prove to be discouraging.

Training in the use of any scale is desirable and invariably improves the rater's judgment. It decreases the variability in range of marks assigned by several such trained judges to the same composition, or of re-ratings made on the same composition by the same judge after different intervals of time. The largest body of available material for practice in marking compositions and comparing them with known values assigned by competent judges is contained in Thorndike's "English Composition—150 Specimens Arranged for Use in Psychological and Educational Experiments," Teachers College, Columbia University, New York City. For other less extensive sources of practice series see the discussion in Chapter II.

FURTHER PROBLEMS

The data which the writer has on hand would both enable her to extend greatly the present investigation, and point to additional problems which ought to be solved in the future for which she has no information. Such problems are:

1. Which scale helps most in determining the strong and weak points of a composition?
2. Which scale is best suited to register small amounts of progress after the application of remedial measures?
3. How much training should a teacher have to be a thoroughly satisfactory judge of composition?
4. Should compositions be marked periodically by a trained marker—one in each school—or should each English teacher receive training in the use of scales?

5. How much improvement in the objectivity of the scale would result from the training of judges to the degree of .5 points average deviation from the true score?

6. What is the relative effect of equal amounts of training upon good and poor judges?

7. What is the effect of a few uniformly extreme judges on the variability of the composite judgment of twenty-five judges?

8. What is the relationship between speed and accuracy in rating?

9. Are English teachers, by the virtue of their experience in the teaching of English, better judges of English composition than non-English teachers?

10. What benefits result from putting a scale into the hands of the pupils?

APPENDIX I

EVIDENCE ON THE INAPPLICABILITY OF THE HARVARD- NEWTON ENGLISH COMPOSITION SCALES TO THE SIXTH GRADE

In Chapter I, the writer promised to present evidence concerning the unsuitability of the Harvard-Newton English Composition Scales for use in the sixth grade. As stated there, Ballou, the author of the scales, definitely states that they are composed of eighth grade compositions, and are to be applied to theme work of that grade. Some persons conducting surveys, however (see end of Chapter IV), have employed them in grades four to seven, inclusive. The data upon which this conclusion is based are:

TABLE 24
HARVARD-NEWTON ENGLISH COMPOSITION SCALES

	Pearson C. V.	S. D.	Per cent of undistributed scores
Before training	12.7082	5.47	
After training	13.8220	5.14	61.6

The smallness of the Pearson coefficient of variation is due to the fact that the sixth grade compositions are so obviously inferior to those in the scale that practically all receive the lowest score which it is possible to give. The same fact, also, accounts for the size of the standard deviation, and its relative constancy, both before and after training. The added information that the per cent of undistributed scores is so high proves the author's claim that this scale is suited for eighth grade (see footnote number 10). Even the lower range is of a quality which is higher than the average performance of sixth grade pupils such as we measured. The character of the scale is such that it might well be used in grades above the eighth—how far above this grade, the writer

has no evidence to say. The reader can see a graphic representation of the same facts by examining Figure II.

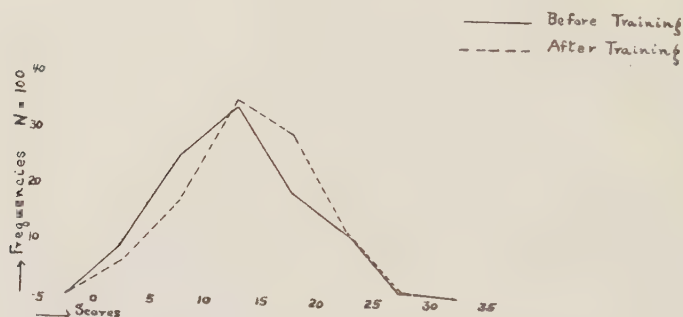


Figure 11 - Harvard-Newton Coefficients of Variation

APPENDIX II

TABLE 25⁴³

IMPROVEMENT AS A RESULT OF TRAINING IN TERMS OF COEFFICIENTS
OF VARIATION GROUPED BY TYPE OF DISCOURSE FOR EACH SCALE

	No. of comps.	H. E. ⁴⁴	H. T.	Leo.	Lew.	Nas.	V. W.	W.(S.)	W.(F.)
Des.	24	3.01	11.58	6.25		1.96	6.23	— .18	— 2.35
Exp.	24	1.50	8.47	5.06		1.84	2.86	1.62	— .85
Nar.	24	2.97	8.35	6.71	3.28	1.84	5.09	.93	— 3.21
Arg.	15	1.56	8.55	3.49	9.00	1.67		2.40	— 3.01
O. L.	13	6.36	7.62	— 2.38	1.64	2.30	5.47	7.57	— 5.49

⁴³ See footnote number 24.

⁴⁴ H. E.—Hudelson English. H. T.—Hudelson Typical. Leo.—Leonard. Lew.—Lewis. Nas.—Nassau. V. W.—Van Wagenen. W. (S.)—Willing (Story Value). W. (F.)—Willing (Form Value).

APPENDIX III

BIBLIOGRAPHY

SCALES AND MANUALS

(The scales which are starred were not used in this study.)

1. Ballou, F. W. "Harvard-Newton Composition Scales." Harvard-Newton Bulletins, Number II, September, 1914, Harvard University, Cambridge, Massachusetts.
2. *Breed, F. S., and Frostic, F. W. "A Scale for Measuring the General Merit of English Composition in the Sixth Grade." University of Chicago, Chicago, Illinois. (Out of print.)
3. *Hillegas, M. B. "A Scale for the Measurement of Quality in English Composition for Young People." Bureau of Publications, Teachers College, Columbia University, New York City, 1913.
4. Hudelson, Earl.
 - (a) "Hudelson English Composition Scale." World Book Company, Yonkers-on-Hudson, New York, 1923.
 - (b) *"Maximal Composition Ability Scale." The Twenty-Second Yearbook of the National Society for the Study of Education, Part I. Public School Publishing Company, Bloomington, Illinois, 1923.
 - (c) "The Hudelson Typical Composition Ability Scale." Public School Publishing Company, Bloomington, Illinois, 1923.
5. Leonard, S. A. "A Scale for the Judgment of Composition Quality Only." English Journal, Vol. 14, December, 1925, pp. 760-775.
6. Lewis, Eugene E. "Lewis English Composition Scales for Measuring Business and Social Correspondence." World Book Company, 1923.
7. Thorndike, Edward L. "Preliminary Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People." Teachers College, Columbia University, New York City.
8. Trabue, M. R. "Nassau County Supplement to the Hillegas Scale." Bureau of Publications, Teachers College, Columbia University, New York City.
9. Van Wagenen, M. J. "Minnesota English Composition Scales" (often known as the Van Wagenen English Composition Scales), World Book Company, Yonkers, New York, 1923.
10. Willing, M. H. "Scale for Measuring Written Composition." Public School Publishing Company, Bloomington, Illinois.

DISCUSSION OF SCALES AND REPORTS OF THEIR DEVELOPMENT AND USE

11. Ashbaugh, E. J. "The Measurement of Language." Journal of Educational Research, June, 1921.
12. Barden, Carrie. "More Experiments with Scales." The English Journal, November, 1924, pp. 648-653.

13. Berman, Samuel. "Standards in Written Composition." Temple University, Philadelphia, 1928.
14. Boston. "The Achievement of Pupils in Letter Writing." Bureau of Educational Investigation and Measurement, Boston Bulletin, Number 16, June, 1918.
15. Briggs, T. H. "English Composition Scales in Use." Teachers College Record, November, 1922, pp. 423-452.
16. Brooks, F. D.
 - (a) "Criteria of Educational Research." School and Society, 1923, December 22, pp. 724-729.
 - (b) "The Relative Accuracy of Ratings Assigned With and Without Use of Drawing Scales." School and Society, Volume 27, April 28, 1928, pp. 518-520.
17. Brown, M. D., and Haggerty, M. E. "The Measurement of Improvement in English Composition." English Journal, Volume 6, October, 1917, pp. 515-527.
18. Brownell, B. "A Test of the Ballou Scale of English Composition." School and Society, Volume 4, December 16, 1916, pp. 938-942.
19. Capps, A. G. "The Functions of Tests and Scales as Found in Recent Educational Literature." Journal of Educational Research, October, 1922, pp. 204-208.
20. Certain, C. C.
 - (a) "Are Your Pupils up to Standard in Composition?" English Journal, June, 1923.
 - (b) "By What Standards Are High School Pupils Promoted in English Composition?" English Journal, Volume 10, June, 1921, pp. 305-315.
 - (c) "Why Not Include Standard Tests in Your Teaching Program This Year?" English Journal, September, 1923, pp. 463-480.
21. Chapman, Harold B. "Organized Research in Education." Ohio State University Studies, Ohio State Press, Columbus, Ohio, 1927.
22. Chapman, J., and Rush, G. P. "The Scientific Measurements of Classroom Products," pp. 131-156. Silver-Burdett Company.
23. Courtis, S. A. "Standard Samples of Known Value for Training Students in the Use of a Scale Sample C." Courtis Standard Research Tests. S. A. Courtis, 246 Eliot St., Detroit, Michigan.
24. Dolch, Edward William. "More Accurate Use of Composition Scales." English Journal, November, 1922, pp. 536-544.
25. Gainsberg, Joseph C. "Fundamental Issues in Evaluating Composition." The Pedagogical Seminary, March, 1924, pp. 55-77.
26. Gilliland, A. R., and Jordan, R. H. "Educational Measurements and the Classroom Teacher," pp. 120-135. The Century Company, 1924.
27. Goodrich, T. V. "Putting Test Scores in a Form to Facilitate Comparison." Journal of Educational Research, December, 1923, pp. 377-388.
28. Gunther, Charles. "My Experience with the Hillegas Scale." English Journal, Volume 8, November, 1919, pp. 535-542.
29. Haggerty, M. E., and Van Wagenen, M. J. "An Investigation into the Amount of Improvement in Ability to Write Composition." Bulletin of the University of Minnesota, 22: 1919. See also the Journal of Educational Psychology, 11: January, 1920, and School and Society 4: November 18, 1916, pp. 761-772.

30. Hosic, James F.
 - (a) "A Summary of Ten Recent Investigations in the Field of English." *Seventeenth Yearbook of the National Society for the Study of Education Part I*, pp. 60-62.
 - (b) "Composition Standards in the Elementary School Arranged to Show the Minimal Performance Essential in the Grades II to VIII." *Seventeenth Yearbook of the N. S. S. E.*, Part I, pp. 46-60.
 - (c) "The Essentials of Composition and Grammar." *Thirteenth Yearbook of the National Society for the Study of Education, Part I*, pp. 90-116. This reference includes also "A Survey of Investigations on Tests and Scales" by the same author.
31. Hudelson, Earl.
 - (a) "English Composition, its Aims, Methods, and Measurement." *Twenty-Second Yearbook of the N. S. S. E.*, Part I, pp. 1-172.
 - (b) "Some Achievements in the Establishment of a Standard for the Measurement of English Composition in Bloomington, Indiana, Schools." *English Journal*, Volume 5, November, 1916, pp. 590-597.
 - (c) "Standards and Measurements in English Composition." *Proceedings of the Second Indiana Conference on Educational Measurements*. Indiana University Press, 1915, Bloomington, Indiana.
 - (d) "The Development and Comparative Value of Composition Scales." *English Journal*, March, 1923, pp. 163-168.
 - (e) "The Effect of Objective Standards upon Composition Teachers' Judgments." *Journal of Educational Research*, December, 1925, pp. 329-340.
32. Hughes, W. Hardin. "Relation of Intelligence to Vocabulary and to Language Training." *English Journal*, October, 1925, pp. 621-626.
33. James, H. W. "A National Survey of the Grading of College Freshman Composition." *English Journal*, Volume 15, October, 1926, pp. 579-587.
34. Jamison, Grace S. "A Study in Correlation of Allied English Abilities." *Journal of Educational Research*, October, 1922, pp. 241-253.
35. Johnston, Nell Bomar. "The Use of Standardized Tests in School Surveys." *Educational Administration and Supervision*, December, 1925, pp. 599-607.
36. Kayfetz, I.
 - (a) "A Critical Study of the Harvard-Newton Composition Scales." *Pedagogical Seminary* 23: September, 1916, pp. 325-347.
 - (b) "A Critical Study of the Hillegas Composition Scale." *Pedagogical Seminary* 21, December, 1914, pp. 559-577.
37. Kelly, F. J. "Teachers' Marks." *Contribution to Education Number 66*, p. 134 ff., 1914. Bureau of Publications, Teachers College, Columbia University, New York City.
38. Klapper, Paul. "Teaching English in Elementary and Junior High Schools." D. Appleton Company, New York, 1925, pp. 174-231.
39. Leonard, S. A. "Building a Scale of Purely Composition Quality." *English Journal*, December, 1925, pp. 760-775.

40. Lewis, Eugene E. "Scales for Measuring Special Types of English Composition." World Book Company, Yonkers, New York.
41. McKinney, I. "An Experiment in Criticism," and "More Comments on Criticism." Illinois Association of Teachers in English, Bulletin 11, October 1, 1916, pp. 1-16 and 10: October 1, 1917, pp. 1-8.
42. Meek, C. S. "Special Report of the Boise Schools." June, 1915.
43. Mirick, G. A. "Progressive Education," pp. 294-300. Houghton Mifflin Company, 1923.
44. Monroe, W. S.
 - (a) "Existing Tests and Standards." Seventeenth Yearbook of the National Society for the Study of Education, Part II, pp. 71-105.
 - (b) "Measuring the Results of Teaching." (Chapter 9.) Houghton Mifflin Company, New York, 1918.
45. Monroe, W. S., DeVoss, J. C., and Kelly, F. J. "Educational Tests and Measurements." Houghton Mifflin Company, 1917.
46. Monroe, W. S., and Others. "Educational Tests and Measurements," pp. 202, 219-221. Houghton Mifflin Company, 1917.
47. Noyes, E. C., and Miller, Edith. "Progress in Standardizing the Measurement of Composition." English Journal, Volume 1, November, 1912, pp. 532-536.
48. Paulu, E. M. "Diagnostic Testing and Remedial Teaching," pp. 165-189. D. C. Heath and Company, 1924.
49. Pooley, Robert. "The Objective Measurement of English Composition." English Journal, June, 1928.
50. Pressey, S. L., and L. C. "Introduction to the Use of Standard Tests," pp. 97-102. World Book Company, Yonkers, New York.
51. Rice, J. M. "The Results of a Test in Language." Forum 35: October, 1903, pp. 269-279. Reprinted in his "Scientific Management in Education"—Hines, Noble and Eldredge, 1913, Philadelphia.
52. Sackett, L. W. "Comparable Measures of Composition." School and Society, 5: February 24, 1917, pp. 233-239.
53. Sangren, Paul U. "Intelligence Tests and the Classification of Students in the Ninth Grade English." Educational Administration and Supervision, December, 1923, pp. 547-553.
54. "Second Report Committee on Minimal Essentials in Elementary School Subjects." Sixteenth Yearbook, Part I, of the National Society for the Study of Education, 1917.
55. Stoddard, W. E. "A Comparison of the Hillegas and Harvard-Newton Scales in English Composition." Pedagogical Seminary 23: December, 1916, pp. 498-501.
56. Theisen, W. W. "Improving Teachers' Estimates of Composition." School and Society 7: pp. 143-150.
57. Thorndike, Edward L.
 - (a) "English Composition—150 Specimens Arranged for Use in Psychological and Educational Experiments." Teachers College, Columbia University, New York City, 1916.
 - (b) "Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition," English Journal, Volume 2, November, 1913, pp. 551-561.
58. Trabue, M. R. "Measuring Results in Education," pp. 362-386. American Book Company, 1924. "Supplementing the Hillegas Scale." Teachers College Record 18: January, 1917, pp. 51-84.

59. Van Wagenen, M. J.
 - (a) "Accuracy with Which English Themes May be Graded with the Use of the English Composition Scales." *School and Society* No. 11: April 10, 1920, pp. 441-450.
 - (b) "The Minnesota Composition Scales: Their Derivation and Reliability." *Educational Administration and Supervision*, 7: December, 1921, pp. 481-499. See also the *Journal of Educational Psychology* 11: November, 1920, pp. 459-573.
60. Ward, C. H. "The Scale Illusion." *English Journal* 6: April, 1917, pp. 221-230.
61. Watson, P. M. "The Harvard-Newton Composition Scale." *Educational Administration and Supervision* I, January, 1915, pp. 57-58.
62. Welch, Roland A. "The Hillegas Scale as a Measure of Composition Work." *English Journal*, Volume 15, October, 1926, pp. 618-620.
63. Willing, M. H.
 - (a) "Individual Diagnosis in Written Composition." *Journal of Educational Research*, February, 1926, pp. 77-89.
 - (b) "The Measurement of Written Composition in Grades Four to Eight." *English Journal* 7: March, 1918, pp. 193-202.
 - (c) "The Measurement of Written Composition in the Public Elementary Schools of Denver, Colorado." Master's Thesis, University of Chicago, Unpublished. See also Denver Survey.
64. Wilson, G. M., and Hoke, K. J. "How to Measure." Chapter 6. Macmillan Company, 1920.

SCHOOL SURVEYS USING COMPOSITION SCALES

65. Report of a Survey of the School System of Butte, Montana. Published by the Board of Education, Butte, Montana, 1914.
66. A Report of the Ohio State School Survey Commission, 1914, pp. 106-156.
67. A Survey of the Port Townsend Public Schools, by Herbert G. Lull. *Bulletin of the University of Washington*, 1915. University Extension Series No. 17, General Series No. 95, pp. 99-102.
68. Report of the Survey of the Public School System of Salt Lake City, Utah. Board of Education, 1915, pp. 138-147.
69. Educational Survey of the Public Schools of Brookline, Massachusetts. Published by the School Committee, Brookline, Massachusetts, 1917, pp. 286-287.
70. School Survey of Denver, Part II, "The Work of the Schools." The School Survey Committee, Denver, Colorado, 1916, pp. 59-64.
71. School Survey of Grand Rapids, Michigan, 1916.
72. Report of a Survey of the School System of St. Paul, Minn., 1917.
73. Report of the Survey of Public Instruction in Nassau County, New York. *University of the State of New York Bulletin*, No. 652, 1917, pp. 146-162.
74. Survey of Janesville, Wisconsin, Public Schools, 1918. Board of Education of the State of Wisconsin.

75. The Gary Public Schools—A General Account—by Flexner and Bachman, p. 88, 98. General Education Board, New York City, 1918.
76. Binghamton School Survey, Binghamton, New York. State Department of Education, 1919.
77. The Baltimore School Survey, pp. 35, 126. Baltimore Board of Commissioners, 1920.
78. Survey of Utica School System, New York State Department of Education, 1919-1920.
79. The Niagara Falls Survey. New York State Department of Education, 1921.
80. The Virginia Public Schools Survey, Part II. "Educational Tests," Volume III, pp. 198-222. World Book Company, 1921.
81. Survey of Lawrence Township, New Jersey, 1922.
82. Survey of the Missouri Elementary Schools. Carnegie Foundation for the Advancement of Teaching, 1922.
83. Survey of Philadelphia Public Schools. State Department of Public Instruction, 1922.
84. Report of the Survey of the Public School System of the Town of Stamford, Connecticut. Institute of Educational Research, Teachers College, Columbia University, New York City, 1923.
85. Report of the Survey of the Schools of Tampa, Florida. The Institute of Educational Research Division of Field Studies Teachers College, Bureau of Publications, Columbia University, 1926, pp. 185-195.

APPENDIX IV

DETAILED INSTRUCTIONS TO JUDGES

SPECIAL DIRECTIONS TO JUDGES ON EACH COMPOSITION SCALE

The author of each of the eight composition scales provided the same with a set of directions. Those sections which were absolutely necessary to know in order to use the scale intelligently were marked in red. In addition to this, a few special directions were given in the case of each scale in order to adapt them to our particular use. A copy of such directions follows.

SPECIAL DIRECTIONS FOR THE HARVARD-NEWTON SCALE

1. Read those parts of pp. 8, 9 and 10 marked in red. (Section 3, page 8. Section 5, page 9. Summary on page 10.)

2. Omit pp. 13-45.

3. Read pp. 45-48. Parts marked in red. (Last paragraph on page 45—a table on page 46, all of pages 47 and 48).

4. Mark all compositions whose code number ends in "a" ("The Funny Things I Saw at the Fair of the Iron Horse") on the "Description Scale"—pp. 49-57.

5. Mark all compositions whose code number ends in "b" ("The Best Way to Get to the Fair of the Iron Horse") on the "Exposition Scale"—pp. 59-68.

On this same scale mark also all compositions whose code number ends in "e" ("A letter to the B. & O. Office, to the director of the pageant, asking for a ticket for a reserved seat in the grand-stand").

6. In the case of letters, disregard the form of the letter and consider only the body of the letter as belonging to the composition. (This is done only when a letter is marked on a non-letter scale.)

7. Mark all compositions whose code number ends in "d" ("A letter to a friend who did not want to go to the Fair of the Iron Horse. Tell how you got him or her to go") on the "Argumentation Scale"—pp. 71-79.

8. Mark all compositions whose code number ends in "c" ("A letter to a friend, telling him or her what a good time you had at the Fair of the Iron Horse") on the "Narration Scale"—pp. 82-93.

9. In spite of the fact that our compositions are only sixth grade compositions and that those of the scale are eighth grade, make no allowances.

10. Beside the mark which you give each composition write the name of the scale on which you marked it. Please assign per cent values instead of letter values.

SPECIAL DIRECTIONS FOR HUDELSON ENGLISH SCALE

1. Read parts marked in red. (Preliminary study of scale.)
2. Omit parts marked "omit."
3. Mark each of the hundred compositions on this scale in the order in which they are arranged.
4. In the case of letters disregard the heading, salutation, and complimentary close, and treat the main body of the letter as a regular composition, remembering that if its title were written it would be "The Fair of the Iron Horse."

SPECIAL DIRECTIONS FOR HUDELSON'S TYPICAL COMPOSITION ABILITY SCALE

First of all, thoroughly familiarize yourself with the scale. Place the scale before you, read the first theme to be scored, compare it with the scale samples one by one, and give it the value of that scale sample which it most nearly resembles in general composition merit. The steps on these scales are fine enough so that for group testing a teacher need not interpolate values. When you have become expert in the use of this scale you will not need to begin with the lowest sample and consult each step in order, but can go at once to that part of the scale which contains samples approximately equal in quality to the theme to be scored. By studying the two or three consecutive samples in that level of the scale you can quickly assign the theme its appropriate score.

Score each of the hundred compositions on this scale. In the case of letters ignore the heading, salutation, and complimentary close, and treat the body of the letter as a regular composition, remembering that if its title were written it would be "The Fair of the Iron Horse."

SPECIAL DIRECTIONS FOR LEWIS ENGLISH COMPOSITION SCALES

1. Read parts marked in red. (Nature of the Scales.)
2. These scales can be used for marking letters only.
3. Mark all letters on the topic, "A letter to the B. & O. Office, to the director of the pageant, asking for a ticket which gave you a reserved seat in the grand-stand," on the first of Lewis' Scales; namely, "A Scale for the Measurement of the Quality of Order Letters."
4. Mark all letters on the topic, "A letter to a friend, telling him or her what a good time you had at the Fair of the Iron Horse," on the second of Lewis' Scales; namely, "A Scale for the Measurement of Narrative Social Letters."
5. Mark all letters on the topic, "A letter to a friend who did not want to go to the pageant. Tell how you got him or her to go," on the third of Lewis' Scales; namely, "A Scale for the Measurement of Expository Social Letters."
6. Apply Lewis' Scales to only 52 compositions of this lot, *i. e.*, letters.

DIRECTIONS FOR LEONARD'S SCALE FOR THE JUDGMENT OF COMPOSITION QUALITY ONLY

1. Read all of this (explicit instructions regarding use of scale) before beginning to mark any compositions.
2. Read the whole scale.
3. The marks to be assigned the compositions are underlined in red.
4. Do *not* interpolate.

DIRECTIONS FOR NASSAU COUNTY SUPPLEMENT TO THE HILLEGAS
SCALE FOR MEASURING THE QUALITY OF ENGLISH COMPOSITIONS

(Directions on scale self-explanatory.)

DIRECTIONS FOR VAN WAGENEN ENGLISH COMPOSITION SCALES

1. Read all parts marked in red. (Nature of the Scales. How to rate compositions.)
2. Omit all parts marked "omit."
3. Mark only 85 of the total number of compositions on this scale in the order in which they are arranged (*i. e.*, the compositions.)
4. Mark all compositions on the topic, "The Best Way to Get to the Fair of the Iron Horse," on the "Exposition Scale." The code number of these compositions ends in "b."
5. Mark all compositions on the topic, "A letter to a friend, telling him or her what a good time you had at the Fair of the Iron Horse," on the "Narration Scale." The code number of these compositions ends in "c."
6. Mark all compositions on the topic, "The Funny Things I Saw at the Fair of the Iron Horse," on the "Description Scale." The code number of these compositions ends in "a."
7. Omit all compositions whose code numbers end in "d."
8. Record marks in proper space on accompanying sheet, giving four marks to each composition: "Thought Content," "Structure," "Mechanics," and "General Merit." (See page 3 of directions.)
9. In front of the code number on the record sheet write the name of the particular one of these scales on which you marked the compositions: thus, "E" for Exposition; "N" for Narration, and "D" for Description.
10. Mark all compositions whose code number ends in "e" (asking for a ticket) on the "Exposition Scale."

DIRECTIONS FOR THE WILLING SCALE FOR MEASURING WRITTEN
COMPOSITIONS

1. Read parts marked in red. (Instructions for rating "Story" and "Form" values.)
2. Mark each of the 100 compositions on this scale in the order in which they are arranged.
3. In the case of letters, disregard the heading, salutation, and complimentary close, and treat the main body of the letter as a regular composition, remembering that if its title were written it would be "The Fair of the Iron Horse."
4. Give two marks to each composition: one for story value, and one for form value.
5. Record these marks in the proper space on the accompanying sheet.
6. For rules of punctuation consult the enclosed mimeographed sheet.
7. You may interpolate if you so desire.

GENERAL DIRECTIONS

1. Get your package of compositions from Mrs. Pearson (Dr. Bamberger's secretary), room 217 Gilman Hall, at any time between nine o'clock in the morning and four-thirty in the afternoon.

2. When finished, return package of compositions to Mrs. Pearson in the envelope exactly as you received it.

3. If in doubt about any detail whatsoever, feel free to call me (Dorothy Speer), Wolfe 6210-W, at any time after 7 P. M.

4. Your package will contain 100 compositions arranged in a certain order. Mark the compositions in the exact order in which they are given to you.

5. Before attempting to use the scale, read the directions printed on the scale. Consult with no one except me, since comparisons will invalidate the experiment.

6. Do not practice marking of compositions provided for practice in some of the scales. With each scale you will be given a blank containing the code numbers of the compositions and the name of the scale by which you are to score them.

7. *Make no marks, corrections or notes of any kind on the mimeographed copies.* If you have any explanation or information to give aside from the mark assigned to the composition and recorded on the sheet opposite the code numbers, write this on a separate sheet of paper, and slip it into the envelope given you.

8. The compositions given you are the actual products of children in the City of Baltimore. The mark you give the composition should be determined by one thing only, and that is the mark given to that sample of scale which most closely resembles in quality the composition which you are marking.

IMPORTANT!

Keep strict account of the number of compositions you mark in an hour. This information concerning the number of compositions a person can mark in an hour, and the increase in his speed of marking with practice, is an important part of the experiment. Write this information on the back of the sheet which contains your marks. Do this as accurately and conscientiously as possible.

9. You will find a good many letters among the 100 compositions. In marking these on any scale except the Lewis Scales consider *only* the body of the letter. The compositions as presented to you in mimeographed form are an attempt to reproduce all the errors, spacing, crossed out and corrected portions of the letter. Consider these when assigning your marks.

10. All the compositions were written on one general topic; namely, "The Fair of the Iron Horse." There were five sub-topics from which the pupil was to choose one. These were as follows:

1. The Funny Things I Saw at the Fair of the Iron Horse. (This calls for description.)

2. The Best Way to Get to the Fair of the Iron Horse. (This calls for exposition.)

3. A letter to a friend telling him or her what a good time you had at the Fair of the Iron Horse. This calls for what Lewis designates as the "Narrative-Social Letter," *i. e.*, it is simple narration in letter form, and can be marked on any scale which does not make a distinction between the discourses (narration, description, exposition, argu-

mentation), or the "Narration Scale" of the Harvard-Newton Scale, and on the "Narration Scale" of the Van Wagenen Scale.

4. A letter to a friend who did not want to go to the Fair of the Iron Horse. Tell how you got him or her to go.

(This calls for argumentation, and must be marked on the scale which Lewis calls the "Expository Social Letters.")

5. A letter to the B. & O. office, to the director of the pageant, asking for a ticket for a reserved seat in the grand-stand. (This calls for an order letter.)

PUNCTUATION RULES

I. *Period:*

1. After all abbreviations.
2. After complete sentences.

II. *Question:*

1. After direct questions.
2. Enclosed in parentheses (?) to indicate a doubtful fact.

III. *Exclamation:*

1. After all exclamatory expressions.
2. To indicate irony.
3. To indicate a strong command.
4. To indicate surprise.

IV. *Commas:*

1. After words in a series.
2. Before the conjunctions of a compound sentence.
3. To set off an introductory phrase.
4. To set off the introductory subordinate clause in a complex sentence.
5. To set off a noun of address.
6. To set off appositives.
7. After the salutation of a friendly letter.
8. To set off a non-restrictive clause.
9. To set off the words "yes" and "no."
10. To set off such introductory expressions as "certainly," "indeed," "by all means."
11. To separate the parts of dates and addresses.

V. *Semi-colon:*

1. In a compound sentence when the conjunction is omitted.
2. Before such words as: accordingly, consequently, however, indeed, nevertheless, still, then, hence, moreover, therefore.

VI. *Colon:*

1. To introduce an enumeration.
2. To precede a long formal quotation.
3. After the salutation of a formal letter.
4. To indicate the time (*e. g.*, 4:30 A. M.).

SAMPLE SCORE SHEET

Sheet No. —

(Name of judge)

How long did it take you to become familiar with the scale?

Record time consumed in marking the hundred compositions at the bottom of the score sheet.

Record marks you assign to the compositions in the proper space below, opposite the appropriate code number.

(Name of the scale)

QWZ-FH-MBS-c	QSQ-ZP-RMR-a
QTX-FM-HM-a	QQU-GB-MF-d
QZW-SC-LF-b	QXV-RF-HM-c
QTW-LN-RLB-a	QUZ-ML-MBS-a
QZV-SC-AMD-c	QZQ-DB-MEF-b
QYX-WD-AMD-b	QQT-FB-RLB-d
QQY-AA-RLB-a	QZZ-JB-MBS-b
QYW-NDB-RLB-b	QWU-EH-HM-c
QQX-JA-MG-e	QYT-LE-MAW-d
QYV-EE-HM-c	QUT-WMC-RMR-c
QXU-EG-MG-c	QYS-EE-RLB-c
QYU-GE-MAW-a	QWY-TH-MBS-d
QRV-CS-RMR-a	QYR-EE-MEF-c
QZY-CC-HM-b	QUS-RMG-LF-b
QUV-HM-RLB-d	QZU-MC-RMR-c
QZX-BC-MF-e	QVQ-FK-AMD-a
QRX-JS-MEF-a	QZT-VC-RMR-b
QXY-DF-MAW-e	QUY-AL-LF-a
QUU-AM-HM-a	QZS-JC-MG-d
QXX-MF-MAW-e	QRU-MT-LF-a
QVR-LK-RMR-a	QWT-RH-MG-b
QXW-LF-MBS-b	QUX-DL-RLB-d
QQV-SA-RLB-a	QWS-JH-LF-b
QTR-WP-MF-d	QVZ-AK-MAW-d
QUQ-AL-RMR-a	QZR-BC-LF-b
QXT-WG-HM-d	QVW-RK-RMR-b
QTV-CN-HM-a	QSR-GS-MEF-c
QRT-EU-HM-b	QYZ-AD-MAW-d
QTU-EOC-RMR-b	QRQ-HS-RLB-e
QSZ-BR-RMR-a	QVV-EK-AMD-a
QWX-TH-LF-c	QVT-AK-AMD-e
QSY-HR-RMR-b	QYY-ED-RMR-c
QOS-FB-LF-b	QRZ-NS-MF-b
QWW-AH-MG-c	QXR-SG-RLB-a
QXQ-DF-HM-c	QVS-JK-RMR-b
QSX-ER-RMR-c	QTZ-DM-EOM-e
QQR-BB-AMD-c	QRY-RS-HM-b
QSW-JR-EOM-e	QWQ-SH-MF-d
QXZ-AF-MAW-e	QRW-RS-RLB-c
QSV-TR-RMR-b	QTY-MM-RMR-e
QUR-TM-RMR-c	QXS-LG-RLB-b
QSU-RR-AMD-a	ZQQ-EZ-HM-a
QVY-BK-MF-e	QQW-LA-HM-c
QWV-LH-HM-c	QQZ-AA-MG-d
QTQ-LM-RMR-a	QRS-FW-AMD-b
QST-MR-EOM-d	QTT-EO-RMR-c
QVX-SK-MAW-c	QUW-JL-RMR-a
QSS-BS-MF-c	QTS-IO-RMR-b
QYQ-KD-MEF-a	QRR-DW-MG-d
QVU-DK-LF-c	QWR-JJ-MBS-e

Record time here: 1st hour 2d hour

Comments on the scale.....

VITA

Dorothy Speer, daughter of the late William and Marie Roth Speer, was born December 15, 1901, in Baltimore, Maryland. She received her entire elementary and secondary education in the public schools of that city. On June 19, 1919, she was graduated from the Eastern High School, at which time she won a Peabody Medal for scholastic achievement, and a scholarship to Goucher College, awarded on the basis of competitive examination. Four years later, June 4, 1923, she received the degree of Bachelor of Arts from Goucher College and was elected to Phi Beta Kappa. Thereupon she taught geography and history in the Chevalier d'Anmour Junior High School, Baltimore, for three years.

From 1926 to 1929, she devoted all her time to graduate work at The Johns Hopkins University. During the second year of her residence, she received a Hopkins Scholarship and election to the Pi Lambda Theta Honorary Fraternity for Women in Education, and during the third she was awarded the George Peabody Scholarship in Education. She is thoroughly interested in all matters pertaining to education, and is a member of the National Society for the Study of Education and the National Education Association.

